Springer Handbook of Spacetime

Abhay Ashtekar, Vesselin Petkov (Eds.)

With 190 Figures and 9 Tables

Springer

Editors
Abhay Ashtekar
Pennsylvania State University
Department of Physics
University Park
PA 16802, USA

Vesselin Petkov
Institute for Foundational Studies Hermann Minkowski
Montreal, Quebec, Canada

ISBN: 978-3-642-41991-1 e-ISBN: 978-3-642-41992-8
DOI 10.1007/978-3-642-41992-8
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2014940760

Production and typesetting: le-tex publishing services GmbH, Leipzig
Senior Manager Springer Handbook: Dr. W. Skolaut, Heidelberg
Typography and layout: schreiberVIS, Seeheim
Illustrations: Hippmann GbR, Schwarzenbruck
Cover design: eStudio Calamar Steinen, Barcelona
Cover production: WMXDesign GmbH, Heidelberg
Printing and binding: Stürtz GmbH, Würzburg

Printed on acid free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

In his *Principia Mathematica*, Isaac Newton formalized the notions of space and time, thereby laying the foundations of a new, revolutionary science. The sweeping success of celestial mechanics firmly established the power of Newton's spacetime paradigm. Soon it became firmly rooted in scientific thought and it gradually came to be an integral part of human consciousness itself. It became commonplace to assume that space is a three-dimensional Euclidean continuum and time flows eternally and uniformly, indifferent to everything else. This view reigned for over 200 years.

However, it was dramatically toppled at the beginning of the twentieth century by the even more revolutionary theories of special and general relativity. First, Albert Einstein taught us that the flow of time is not so indifferent after all; time intervals depend on the state of motion of the observer. Hermann Minkowski completed this scenario by showing us that space and time, in fact, fuse together to form a genuinely four-dimensional spacetime continuum. As he put it,

> *the whole world presents itself as resolved into such worldlines, and I want to say in advance, that in my understanding the laws of physics can find their most complete expression as interrelations between these worldlines.*

This fusion served as the point of departure for Einstein's discovery of general relativity. In this theory, spacetime geometry is no longer flat. Its curvature encodes the gravitational field. Spatial distances and time intervals between events are replaced by the proper time elapsed between them along worldlines of observers. This duration is sensitive not only to the motion of those observers but also to the gravitational field in the region. Space and time are no longer inert, background entities, a canvas on which the dynamics of particles and fields is painted. Spacetime *itself* is now dynamical, an active player in the drama of evolution. This new conceptual framework is truly compelling. As Hermann Weyl said,

> *It is as if a wall that separated us from truth has collapsed. Wider expanses and greater truths are now exposed to the searching eye of knowledge, regions of which we had not even a pre-sentiment.*

Soon after his discovery of general relativity, Einstein wrote to Arnold Sommerfield:

> *Of the general theory of relativity, you will be convinced, once you have studied it. Therefore, I am not going to defend it with a single word.*

This Springer Handbook of Spacetime is dedicated to the ground-breaking paradigm shifts embodied in the two relativity theories and describes in detail the profound reshaping of the physical sciences that they ushered in. In a single volume it includes chapters on the foundations, the underlying mathematics, physical and astrophysical implications, experimental evidence and cosmological predictions, as well as chapters on efforts to unify general relativity and quantum physics. The presentation is at an introductory level in that each chapter provides a bird's-eye view of a sub-area in which notable advances have occurred, especially in the past 30 years. Therefore, the Handbook can be used as a ready reference by researchers in a wide variety of fields, not only by specialists in relativity but also by researchers in related areas that either grew out of, or are deeply influenced by, the two relativity theories: cosmology, astronomy and astrophysics, high-energy physics, quantum field theory, mathematics, and the philosophy of science. It should also serve as a valuable resource for graduate students and young researchers entering these areas, and for instructors who teach courses on these subjects.

The Springer Handbook of Spacetime is divided into six parts. The first part deals with the historical origins of the spacetime notion that emerged from special relativity and introduces the basic ideas of special and general relativity. It ends with an emphasis on the intrinsic link between physics and spacetime geometry revealed by the two relativity theories. The second part is devoted to a number of foundational issues, most of which are concerned with the nature of time and gravity. This part also discusses some subtle issues in special and general relativity. The third part introduces the reader to mathematical structures that have served as powerful tools to unravel numerous implications of the two relativity theories. Here, the emphasis is on theoretical frameworks that are widely used in the contemporary research on spacetime structures, and on the

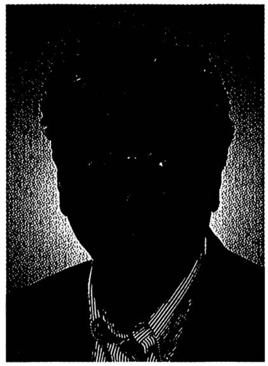
qualitatively new results that have emerged naturally. Because they are unrelated to the initial motivations used by Einstein, these unexpected advances bring out the amazing depth of general relativity and, more generally, the richness of the interplay between physics and mathematics. The fourth and the fifth parts summarize the observational status of the two relativity theories and the deep influence general relativity has had on our understanding of the cosmos as a whole. Here, one finds another amazing synergy, namely that between advanced technology and predictions of general relativity. One cannot be but deeply impressed by the fact that not only is the theory exceptional in its aesthetic beauty – its supreme conceptual economy and mathematical elegance – but it has also withstood some of the most stringent and imaginative observational tests to which any physical theory has been subjected. The sixth and final part illustrates various approaches to the unification of general relativity and quantum physics. They provide a flavor of the new science that could lead us to the next paradigm shift, taking us well beyond our present notion of spacetime.

This Springer Handbook is the outcome of the dedicated effort and commitment of many individuals. Authors accepted the difficult task of pitching their chapters at a level that is suitable for beginning researchers in the field and readily incorporated suggestions for improvements made by the referees. Numerous referees sent very detailed and helpful comments on manuscripts. Angela Lahee coordinated a smooth and delightful collaboration with Springer. This project could never have been completed without the generous support of all these individuals. We are grateful to them all. This work was supported in part by the NSF grant PHY-1205388 and the Eberly Research Funds of Penn State.

June 2013

Abhay Ashtekar — University Park, PA, USA
Vesselin Petkov — Montreal, Quebec, Canada

About the Editors



Prof. Abhay Ashtekar received his PhD from the University of Chicago and was awarded Doctor Rerum Naturalium Honoris Causa by the Friedrich-Schiller Universität, Jena, Germany and by Université de la Mediterranée, Aix-Marseille, France. Currently, he is the Director of the Institute for Gravitation and the Cosmos at Pennsylvania State University where he also holds the Eberly Chair in Physics. He is a Fellow of the American Association for the Advancement of Science, a Honorary Fellow of the Indian Academy of Sciences, and is a past President of the International Society for General Relativity and Gravitation.



Vesselin Petkov received a graduate degree in physics from Sofia University, a doctorate in philosophy from the Institute for Philosophical Research of the Bulgarian Academy of Sciences, and a doctorate in physics from Concordia University in Montreal. He taught at Sofia University and Concordia University, and also had a stint at the Physics Department of the Johannes Kepler University of Linz, Austria, before coming to Montreal in 1990. He is one of the founding members of the Institute for Foundational Studies *Hermann Minkowski.*

About the Authors

Ivan Agullo
University of Cambridge
Department of Applied Mathematics and Theoretical Physics
Wilberforce Road
Cambridge, CB3 OWA, UK
and
Louisiana State University
Department of Physics & Astronomy
Tower Dr.
Baton Rouge, LA 70803-4001, USA
e-mail: *i.agullorodenas@damtp.cam.ac.uk*

Jan Ambjørn
Copenhagen University
The Niels Bohr Institute
Blegdamsvej 17
2100, Copenhagen, Denmark
and
Radboud University Nijmegen
Institute for Mathematics, Astrophysics and Particle Physics (IMAPP)
Heyendaalseweg 135
6500 GL, Nijmegen, Netherlands
e-mail: *ambjorn@nbi.dk*

Neil Ashby
NIST
Time and Frequency Division
325 Broadway, Div 688
Boulder, CO 80305, USA
e-mail: *ashby@boulder.nist.gov*

Beverly K. Berger
2131 Chateau PL
Livermore, CA 94550, USA
e-mail: *beverlyberger@me.com*

Orfeu Bertolami
Universidade do Porto
Faculdade de Ciências, Departamento de Física e Astronomia
Rua do Campo Alegre 687
4169-007, Porto, Portugal
e-mail: *orfeu.bertolami@fc.up.pt*

Robert T. Bluhm
Colby College
Department of Physics and Astronomy
Waterville, ME 04901, USA
e-mail: *rtbluhm@colby.edu*

Sergio del Campo
Pontificia Universidad Catolica de Valparaiso
Av. Universidad 330, Curauma
Valparaíso, Chile
e-mail: *sdelcamp@ucv.cl*

Alejandro Corichi
National Autonomous University of Mexico (UNAM)
Centro de Ciencias Matematicas, Quantum Gravity Group, UNAM Campus Morelia
A. Postal 61-3
58089, Morelia, Michoacan, Mexico

Sergio Dain
Universidad Nacional de Córdoba
Facultad de Matemática, Astronomía y Física
Medina Allende y Haya de la Torre
5000, Ciudad Universitaria, Córdoba, Argentina
e-mail: *dain@famaf.unc.edu.ar*

Diako Darian
Tvetenveien 215
0675, Oslo, Norway
e-mail: *diako.darian@gmail.com*

Dennis Dieks
Utrecht University
History and Foundations of Science
3508 TA, Utrecht, Netherlands
e-mail: *d.dieks@uu.nl*

George F.R. Ellis
University of Cape Town
Department of Mathematics
7701, Rondebosch, Cape Town, South Africa
e-mail: *george.ellis@uct.ac.za*

Jonathan S. Engle
Florida Atlantic University
Department of Physics
777 Glades Road
Boca Raton, FL 33431-0991, USA
e-mail: *jonathan.engle@fau.edu*

Rafael Ferraro
Instituto de Astronomía y Física del Espacio
Casilla de Correo 67, Sucursal 28
1428, Buenos Aires, Argentina
and
Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales,
Departamento de Física
Ciudad Universitaria, Pabellón I
1428, Buenos Aires, Argentina
e-mail: *ferraro@iafe.uba.ar*

Sebastian Fischetti
University of California Santa Barbara
Department of Physics
Broida Hall
Santa Barbara, CA 93106, USA
e-mail: *sfischet@physics.ucsb.edu*

Maurizio Gasperini
University of Bari
Department of Physics
Via G. Amendola 173
70126, Bari, Italy
e-mail: *gasperini@ba.infn.it*

Robert Geroch
The University of Chicago
Department of Physics
5720 S. Ellis Ave
Chicago, 60637, USA
e-mail: *geroch@uchicago.edu*

Domenico Giulini
Leibniz Universität Hannover
Institut für Theoretische Physik
Appelstrasse 2
30167, Hannover, Germany
e-mail: *giulini@itp.uni-hannover.de*

Andrzej Görlich
Copenhagen University
The Niels Bohr Institute
Blegdamsvej 17
2100, Copenhagen, Denmark
and
Jagellonian University
Marian Smoluchowski Institute of Physics,
Department of Theory of Complex Systems
ul. Reymonta 4
30-059, Kraków, Poland
e-mail: *atg@th.if.uj.edu.pl*

Øyvind Grøn
Oslo and Akershus University College of Applied Sciences
0130, Oslo, Norway
and
University of Oslo
Institute of Physics
0316, Oslo, Norway
e-mail: *oyvind.gron@hioa.no*

Graham S. Hall
University of Aberdeen
Institute of Mathematics
Aberdeen, AB24 3UE, UK
e-mail: *g.hall@abdn.ac.uk*

James Isenberg
University of Oregon
Department of Mathematics
Eugene, OR 97403, USA
e-mail: *isenberg@uoregon.edu*

Pankaj S. Joshi
Tata Institute of Fundamental Research
Homi Bhabha Road
400005, Mumbai, India
e-mail: *psj@tifr.res.in*

Jerzy Jurkiewicz
Jagellonian University
Marian Smoluchowski Institute of Physics,
Department of Theory of Complex Systems
ul. Reymonta 4
30-059, Kraków, Poland
e-mail: *jurkiewicz@th.if.uj.edu.pl*

William Kelly
University of California Santa Barbara
Department of Physics
Broida Hall
Santa Barbara, CA 93106, USA
e-mail: *wkelly@physics.ucsb.edu*

Claus Kiefer
University of Cologne
Institute for Theoretical Physics
Zülpicher Strasse 77
50937, Köln, Germany
e-mail: *kiefer@thp.uni-koeln.de*

Badri Krishnan
Max Planck Institute for Gravitational Physics
Observational Relativity and Cosmology
Callinstr. 38
30167, Hannover, Germany
e-mail: *badri.krishnan@aei.mpg.de*

Renate Loll
Radboud University Nijmegen
Institute for Mathematics, Astrophysics and Particle Physics (IMAPP)
Heyendaalseweg 135
6500 GL, Nijmegen, Netherlands
e-mail: *r.loll@science.ru.nl*

Stephen N. Lyle
La Coumo d'Andebu
09240, Alzen, France
e-mail: *stephen.n.lyle@gmail.com*

Donald Marolf
University of California Santa Barbara
Department of Physics
Broida Hall
Santa Barbara, CA 93106, USA
e-mail: *marolf@physics.ucsb.edu*

Thanu Padmanabhan
IUCAA
Pune University Campus
411 007, Pune, India
e-mail: *paddy@iucaa.ernet.in*

Jorge Páramos
Universidade do Porto
Faculdade de Ciências, Departamento de Física e Astronomia
Rua do Campo Alegre 687
4169-007, Porto, Portugal
e-mail: *jorge.paramos@fc.up.pt*

José G. Pereira
Universidade Estadual Paulista
Instituto de Física Teórica
01140-070, São Paulo, Brazil
e-mail: *jpereira@ift.unesp.br*

Vesselin Petkov
Institute for Foundational Studies
Hermann Minkowski
Montreal, Quebec, Canada
e-mail: *vpetkov@minkowskiinstitute.org*

Goswami Rituparno
University of KwaZulu-Natal
School of Mathematics Statistics and Computer Science, Westville Campus
4000, Durban, South Africa
e-mail: *vitasta9@gmail.com*

Carlo Rovelli
Aix-Marseille University
Centre de Physique Théorique de Luminy
Luminy
13288, Marseille, France
e-mail: *rovelli@cpt.univ-mrs.fr*

Lewis Ryder
University of Kent
School of Physical Sciences, Ingram Building
Kent
Canterbury, CT2 7NH, UK
e-mail: *l.h.ryder@kent.ac.uk*

Hanno Sahlmann
Friedrich-Alexander University
Erlangen-Nürnberg
Department of Physics
Staudtstr. 7
91058, Erlangen, Germany
e-mail: *hanno.sahlmann@gravity.fau.de*

B. Suryanarayana Sathyaprakash
Cardiff University
School of Physics and Astronomy
5 The Parade
Cardiff, CF24 3AA, UK
e-mail: *b.sathyaprakash@astro.cf.ac.uk*

Tarun Souradeep
Inter-University Centre for Astronomy and Astrophysics (IUCAA)
Ganeshkhind
411007, Pune, India
e-mail: *tarun@iucaa.ernet.in*

Norbert Straumann
University of Zurich
Institute for Theoretical Physics
Winterthurerstrasse 190
8057, Zurich, Switzerland
e-mail: *norbert.straumann@gmail.com*

Chris Van Den Broeck
National Institute for Subatomic Physics (Nikhef)
Department of Gravitational Physics
Science Park 105
1098 XG, Amsterdam, Netherlands
e-mail: *vdbroeck@nikhef.nl*

Scott Walter
University of Lorraine
LHSP-Archives Henri-Poincaré (CNRS, UMR 7117)
91 avenue de la Libération, BP 454
54001, Nancy, France

David Wands
University of Portsmouth
Institute of Cosmology and Gravitation
Burnaby Road
Portsmouth, P01 3FX, UK
e-mail: *david.wands@port.ac.uk*

Paul S. Wesson
University of Waterloo
Department of Physics and Astronomy
Waterloo, Ontario, N2L 4G1, Canada
e-mail: *psw.papers@yahoo.ca*

Nick M. J. Woodhouse
University of Oxford and Clay Mathematics Institute
Mathematical Institute, Andrew Wiles Building, Radcliffe Observatory Quarter
Woodstock Road
Oxford, UK
e-mail: *nick.woodhouse@maths.ox.ac.uk*

H. Dieter Zeh
Universität Heidelberg
Institut für Theoretische Physik
Philosophenweg 19
69120, Heidelberg, Germany
e-mail: *zeh@uni-heidelberg.de*

Contents

List of Abbreviations

Symbols

ΛCDM	Lambda-cold dark matter
1-D	one-dimensional
2dFGRS	2def Galaxy Redshift Survey

A

AAdS	asymptotically AdS
ACT	Atacama Cosmology Telescope
ADM	Arnowitt, Deser, Misner
AdS	anti-de Sitter
AlAdS	asymptotically locally AdS
aLIGO	advanced LIGO
AO	accelerating observer
ATHENA	Apparatus for High Precision Experiments on Neutral Antimatter
ATRAP	Antihydrogen trap
AVTD	asymptotically velocity term dominated

B

BBH	binary black hole
BBN	big-bang nucleosynthesis
BD	Bunch–Davies
BF	Breitenlohner–Freedman
BICEP	background imaging of cosmic extragalactic polarization
BIPM	International Bureau of Weights and Measures
BKL	Belinski, Khalatnikov, Lifshitz
BNS	binary neutron star
BTZ	Bañados-Teitelboim–Zanelli

C

CBU	crystallizing block universe
CDT	causal dynamical triangulation
CERN	European Organization for Nuclear Research
CFT	conformal field theory
CMB	cosmic microwave background
CMBR	cosmic microwave background radiation
CMC	constant mean curvature
COBE	Cosmic Background Explorer
COW	Colella, Overhauser, and Werner
CPLEAR	charge parity low-energy antiproton ring
CPT	charge, parity, time
CTS	conformal thin sandwich
CW	continuous wave

D

DASI	Degree Angular Scale Interferometer
DECIGO	Decihertz Interferometer Gravitational-Wave Observatory
DH	dynamical horizon
DIRBE	diffuse infrared background experiment
DMR	differential microwave radiometer
DT	dynamical triangulation

E

EBU	emergent block universe
EBU	evolving block universe
ECEF	earth-fixed reference frame
ECI	earth-centered, locally inertial
eLISA	evolved Laser Interferometer Space Antenna
EMRI	extreme mass ratio inspiral
EP	equivalence principle
EP	Einstein equivalence principle
EPRL	Engle-Pereira-Rovelli-Livine
EPRL	Engle-Pereira-CR-Livine
ESA	European Space Agency
ESU	Einstein static universe
ET	Einstein telescope

F

FFF	freely falling frame
FIRAS	far-infrared absolute spectrophotometer
FK	Freidel–Krasnov
FLRW	Friedmann–Lemaître–Robertson–Walker
FOTH	future-outer-trapping horizon
FPTL	future-pointing timelike
FRW	Friedmann–Robertson–Walker
FRWL	Friedmann, Robertson, Walker, and Lemaître

FS	Friedman–Scarr
FW	Fermi–Walker

G

GLONASS	globalnaya navigatsionnaya sputnikovaya sistema
GNSS	global navigation satellite system
GPS	global positioning system
GR	general relativity
GRB	gamma-ray burst
GTRF	Galileo terrestrial reference frame
GUA	generalized uniform acceleration
GUT	grand unification theory
GW	gravitational wave
GZK	Greisen–Zatsepin–Kuzmin

H

H–J	Hamilton–Jacobi
HBBM	hot big bang model
HILV	LIGO-Hanford, LIGO-India, LIGO-Livingston, Virgo
HiRes	high resolution fly's eye
HLV	advanced LIGO detectors plus advanced Virgo
HLVJ	advanced LIGO detectors plus advanced Virgo plus KAGRA
HLVJI	five-detector network plus IndIGO
HOS	hyperplane of simultaneity
HST	Hubble space telescope

I

ICIF	instantaneously comoving inertial frame
ICRF	international celestial reference frame
IERS	International Earth Rotation Service
iLIGO	initial LIGO
IR	infrared
ISS	International Space Station
ITRF	International Terrestrial Reference Frame
ITRS	International Terrestrial Reference System

K

KAGRA	Kamioka Gravitational Wave Detector
KKL	Kamiński, Kisielowski, and Lewandowski
KS	Kostelecký–Samuel
KVF	Killing vector field

L

LAD	Lorentz–Abraham–Dirac
LCBY	Lichnerowicz, Choquet-Bruhat, York
LHC	large hadron collider
LIGO	Laser Interferometer Gravitational-Wave Observatory
LISA	Laser Interferometer Space Antenna
LIVE	Lorentz invariant vacuum energy
LL	Landau–Lifschitz
LLI	local Lorentz invariance
LLR	lunar laser ranging
LMXB	low-mass x-ray binary
LPI	local position invariance
LQC	loop quantum cosmology
LQG	loop quantum gravity
LRF	local Rindler frame
LSF	logarithmic scale factor
LSS	large-scale structure

M

MCP	method of consistent potentials
MEO	medium earth orbit
MIS	Müller–Israel–Stewart
mSME	minimal standard model extension
MSS	minisuperspace
MTT	marginally trapped tube

N

NG	Nambu–Goldstone
NSBH	a neutron star and a black hole
NUT	Newman, Unti, Tamburino

O

OCDM	open cold dark matter
OHD	observed Hubble-parameter dataset

P

PDE	partial differential equation
PGL	projective general linear
PI	principle investigator
PN	post-Newtonian
ppE	parameterized post-Einsteinian

PPN	parameterized post-Newtonian
PTA	pulsar timing array

Q

QCD	quantum chromodynamics
QED	quantum electrodynamics
QFT	quantum field theory
QNM	quasi-normal mode

R

RTG	radiothermal generator

S

SCC	strong cosmic censorship
SCDM	standard cold dark matter
SE	semi-Euclidean
SEP	strong equivalence principle
SHO	simple harmonic oscillator
SKA	Square-Kilometer Array
SM	standard model
SMBBH	supermassive black hole binar
SME	standard model extension
SNAG	Stone–Naimark–Ambrose–Godement
SNR	signal-to-noise ratio
SPT	South Pole Telescope
SR	special relativity
SR	special theory of relativity
SUGRA	supergravity
SV	satellite vehicle

T

TAI	international atomic time
TDSE	time-dependent Schrödinger equation
TIGER	Test Infrastructure for GEneral Relativity
TISE	time-independent Schrödinger equation
TT	transverse traceless
TT	terrestrial time
TUA	translational uniform acceleration

U

USNO	U.S. Naval Observatory
UTC	universal coordinated time
UV	ultraviolet

V

vev	vacuum expectation value
VIRGO	variability irradiance and gravity oscillations
VTD	velocity term dominated

W

WAAS	wide area augmentation system
WDB	white dwarf binary
WDW	Wheeler–De Witt
WEP	weak equivalence principle
WKB	Wentzel–Kramers–Brillouin
WMAP	Wilkinson microwave anisotropy probe

Part A Introduc

Part A Introduction to Spacetime Structure

1. From Æther Theory to Special Relativity

Rafael Ferraro

At the end of the nineteenth century light was regarded as an electromagnetic wave propagating in a material medium called *ether*. The speed c appearing in Maxwell's wave equations was the speed of light with respect to the ether. Therefore, according to the Galilean addition of velocities, the speed of light in the laboratory would differ from c. The measure of such a difference would reveal the motion of the laboratory (the Earth) relative to the ether (a sort of *absolute motion*). However, the Earth's absolute motion was never evidenced.

Galilean addition of velocities is based on the assumption that lengths and time intervals are *invariant* (independent of the state of motion). This way of thinking about the spacetime emanates from our daily experience and lies at the heart of Newton's classical mechanics. Nevertheless, in 1905 Einstein defied Galilean addition of velocities by postulating that light travels at the same speed c in any inertial frame. In doing so, Einstein extended the *principle of relativity* to the electromagnetic phenomena described by Maxwell's laws. In Einstein's special relativity ether does not exist and absolute motion is devoid of meaning. The invariance of the speed of light forced the replacement of Galilean transformations with Lorentz transformations. Thus, relativistic length contractions and time dilations entered into our understanding of spacetime. Newtonian mechanics had to be reformulated, which led to the discovery of the mass-energy equivalence.

1.1 Space and Time in Classical Mechanics

Until 1915, when Einstein's general relativity radically changed our way of thinking, *spacetime* was regarded as the immutable scenery where physical phenomena take place. The laws of mechanics, which describe the motion of a particle subject to interactions, were written to work in this immutable scenery. The form of these laws strongly depends on the properties attributed to the spacetime. Classical mechanics relies on the assumption that distances and time intervals are invariant. This assumption, which seems to be in agreement with our daily experience, leads to the Galilean addition of velocities, which prevents invariant velocities in classical mechanics.

1.1.1 Invariance of Distances and Time Intervals

Classical mechanics – the science of mechanics founded by Newton – considered that space is properly described by Euclid's plane geometry. Then there exist the Cartesian coordinates (x, y, z), so the distance d between two points placed at (x_1,y_1,z_1) and (x_2,y_2,z_2) can be computed by means of the Pythagorean formula

$$d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2 \,. \tag{1.1}$$

In addition, classical mechanics regards distances and time intervals as *invariant* quantities. Let us explain the meaning of this property with an example from daily life concerning the invariance of time intervals. Mario frequently flies from Buenos Aires to Madrid; he knows that the journey lasts 12 h as measured by his watch. This time, Mario wants his friend Manuel to pick him up at Madrid airport. When the flight is about to depart, Mario calls Manuel, who tells him that it is 9 a.m. in Madrid. Then Mario asks Manuel to wait for him at 9 p.m. at Madrid airport, just when the plane will land. This way of arranging a meeting assumes that the time elapses in the same way both in the plane and on Earth. Of course, it seems to be a good assumption because it works effectively in our daily life. We call a magnitude "invariant" if it has the same value in different frames in relative motion (as the plane and the Earth in the previous example). Classical mechanics considers that not only time intervals are invariant but distances too. In particular, the length of a body is assumed to be independent of its state of motion. We can *verify* this assumption in our daily life. For instance, we can measure a train by spreading a tape measure along the train. The so obtained length will seem to agree with a measure performed along the rail while the train is traveling. Notice that measuring the length of a moving body requires some care; the length is the distance between *simultaneous* positions of the ends of the body. In the case of the train, we can imagine that the rail is provided with sensors detecting the stretch of rail that the train takes up at each instant. We can then determine the length of such a stretch of rail by means of a tape measure identical to the one used on the train.

The invariance of distances and time intervals is a property that is supported by our daily experience. It could be said that space and time look like separated concepts to us, and this separation seems not to be affected by the choice of frame. This somehow naive way of regarding space and time is a key piece in the construction of classical mechanics. However, to what extent should we be confident of our daily experience? Does our daily experience cover the entire range of phenomena, or it is rather limited? Let us use a familiar example to explain what we mean: we could well believe that the Earth's surface is flat if just a little portion of it were accessible to us. However, we realize that the Earth's surface is nearly spherical by considering it at larger scales. In this example, the scale should be comparable to the radius of the globe. In the case of the behavior of distances and time intervals under changes of frame, the scale in question is the relative velocity V between the frames. How can we be sure that the invariance of distances and time intervals is nothing but an appearance caused by the narrow range of relative velocities V covered by our daily experience? As we will explain in Sect. 1.4, Einstein's special relativity of 1905 abolished the invariance of distances and time intervals on the basis of new physics developed in the second half of the nineteenth century.

1.1.2 Addition of Velocities

Velocities are not invariant in classical mechanics. Let us consider the motion of a passenger along a train traveling on the rails at $100\,\mathrm{m\,s^{-1}}$. The train and the Earth are two possible frames to describe the motion of the passenger; they are in relative motion at $V = 100\,\mathrm{m\,s^{-1}}$. It is evident that the velocity of the passenger is different in each frame. For instance, the passenger could be at rest on the train, and thus moving at $100\,\mathrm{m\,s^{-1}}$ with respect to the Earth. If the passenger walks for-

ward at a velocity of $u' = 1\ \mathrm{m\,s^{-1}}$, then he/she advances 1 m on the train (as measured by a tape measure fixed to the train) each 1 s (as measured by a clock fixed to the train). Now, how fast does he/she move with respect to the Earth? The answer to this simple question depends on the properties of distances and time intervals under change of frame. Since classical mechanics assumes that distances and time intervals are invariant, we can state that the passenger advances 1 m on the train each 1 s as measured by a clock and a tape measure fixed *to the Earth* (but otherwise identical to those fixed to the train). Besides, in this frame also the train advances at the rate of 100 m each second. Then, the passenger displaces 101 m each second. Thus his/her velocity in the frame fixed to the Earth is $u = 101\ \mathrm{m\,s^{-1}} = u' + V$. This *addition of velocities* is a direct consequence of the classical invariance of distances and time intervals. It means that velocities are not invariant in classical mechanics; they always change by the addition of V. On the contrary, Einstein's special relativity will rebuild our way of regarding space and time by postulating an invariant velocity: the speed of light c ($c = 299\,792\,458\ \mathrm{m\,s^{-1}}$). The postulate of invariance of the speed of light implies the abandonment of our belief in the invariance of distances and time intervals, so strongly rooted in our daily experience. Therefore, deep theoretical and experimental reasons should be alleged to propose such a drastic change of mind. In fact, the idea of invariance of the speed of light is theoretically linked to Maxwell's electromagnetism and the principle of relativity, as will be analyzed in Sect. 1.3. Besides, at the end of the nineteenth century there was enough experimental evidence about the invariance of c. However, those experimental results were not correctly interpreted until special relativity came on stage.

The existence of an invariant speed provides us with a scale of reference to understand why distances and time intervals seem to be invariant in our daily life: according to special relativity, distances and time intervals behave as if they were invariant when the compared frames (the train, the plane, the Earth, etc.) move with a relative velocity $V \ll c$. So, it is just an appearance; like the Earth's surface, which seems to be flat if it is only explored in distances much smaller than the radius of the globe.

1.1.3 Coordinate Transformations

An *event* is a point in the spacetime. It represents a place in space and an instant of time; it is a "here and now". An event is characterized by four coordinates; we will use three Cartesian coordinates x, y, z to localize the place of the event plus its corresponding time coordinate t. Cartesian coordinates are distances measured with rules along the Cartesian axes of the frame. The coordinate t is measured by clocks counting the time from an instant conventionally chosen as the time origin.

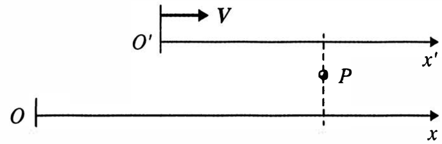


Fig. 1.1 Frames S and S' moving at the relative velocity V

Figure 1.1 shows two frames S and S' in relative motion; the x and x' axes have the direction of the relative velocity V. By comparing distances in the frame S, we can state

$$d_{OP}|_S = d_{OO'}|_S + d_{O'P}|_S\,. \qquad (1.2)$$

In the frame S, the distance between O – the coordinate origin of S – and the place P is the x coordinate of P: $d_{OP}|_S = x$. On the other hand, the distance between the origins O and O' increases with time; if V is constant and the time t in S is chosen to be zero when both origins coincide, then $d_{OO'}|_S = Vt$. Thus

$$d_{O'P}|_S = x - Vt\,. \qquad (1.3)$$

We are not allowed to replace the left member with x', since $x' = d_{O'P}|_{S'}$. Classical mechanics, however, assumes that distances have the same value in all the frames. Thus, we obtain the Galilean transformations.

Galilean Transformations

$$x' = x - Vt\,, \qquad (1.4a)$$

$$y' = y, \qquad (1.4b)$$

$$z' = z. \qquad (1.4c)$$

We have added the transformations of the Cartesian coordinates y, z transversal to the relative motion of the frames. These are distances between a given place and the straight line shared by the x and x' axes; according to the classical invariance of distances, they are equal in S and S'.

The classical transformations of the coordinates of an event is completed by considering the invariance

of time intervals; so we state that $t' = t$ (we choose a common time origin for S and S'). Remarkably, the relation $t' = t$ also results from the transformation (1.4a) with the help of a simple physical argument: as frames S and S' are on an equal footing, the respective inverse transformation should look like (1.4a) except for the sign of V (if S' moves towards increasing values of x in S, then S moves towards decreasing values of x' in S'; thus the relative velocity changes sign). Therefore,

$$x = x' + Vt' . \tag{1.5}$$

Then, by adding (1.4a) and (1.5) one obtains

$$t' = t . \tag{1.6}$$

Galilean Addition of Velocities

A moving particle traces a succession of events in spacetime. This *world-line* can be described by equations $x(t)$, $y(t)$, $z(t)$, which are summarized in a sole vector equation for the position vector $\boldsymbol{r}(t)$. According to Galilean transformations (1.4), the position vector transforms as

$$\boldsymbol{r}'(t) = \boldsymbol{r}(t) - \boldsymbol{V}t , \tag{1.7}$$

where the invariance of time, $t' = t$, has also been used. Differentiating (1.7) results in the *Galilean addition of velocities*, i. e., the relation between the velocities of the particle in two different frames due to the movement composition with the relative translation between both frames

$$\boldsymbol{u}'(t) = \boldsymbol{u}(t) - \boldsymbol{V} . \tag{1.8}$$

Velocities are not invariant under Galilean transformations. However, the relative velocity between two particles is invariant

$$\boldsymbol{u}_2'(t) - \boldsymbol{u}_1'(t) = \boldsymbol{u}_2(t) - \boldsymbol{u}_1(t) . \tag{1.9}$$

Galilean Invariance of the Acceleration

Since $\boldsymbol{V}$ is uniform, the differentiation of (1.8) yields the Galilean invariance of the acceleration

$$\boldsymbol{a}'(t) = \boldsymbol{a}(t) . \tag{1.10}$$

1.2 Relativity in Classical Mechanics

Mechanics describes the motion of interacting particles by means of equations governing the particle world-lines. These equations of motion, together with the initial conditions, yield the coordinates of particles as functions of time: $x(t)$, $y(t)$, and $z(t)$. To write the equations of motion we combine the laws of dynamics with the laws of the interactions. Both types of laws must have the same form in all the inertial frames. This is the principle of relativity in mechanics, which expresses that all the inertial frames are on an equal footing. However, whether or not a given law consummates the principle of relativity is a matter depending on the properties attributed to space and time.

1.2.1 Newton's Laws of Dynamics

Newton constructed the dynamics on the basis of three laws [1.1]:

- *First law* (*principle of inertia*): free particles move with constant velocity (they describe straight world-lines in spacetime).
- *Second law*: a particle acted by a force acquires an acceleration that is proportional to the force

  $$\boldsymbol{F} = m\boldsymbol{a} . \tag{1.11}$$

  The proportionality constant m is a property of the particle called *mass*. In terms of the *momentum* $\boldsymbol{p} \equiv m\boldsymbol{u}$, the law reads $\boldsymbol{F} = \mathrm{d}\boldsymbol{p}/\mathrm{d}t$.
- *Third law* (*action-reaction principle*): two particles interact by simultaneously exerting each other equal and opposite forces.

The first law is a particular case of the second law (the case $\boldsymbol{F} = 0$); it establishes the tendency to perdurability as the main feature of motion (as was envisaged by *Galileo* [1.2], *Gassendi* [1.3], and *Descartes* [1.4], in opposition to Aristotelian thought). On the other hand, the second law becomes the particle equation of motion, once the force is given as a function of $\boldsymbol{r}$, $\boldsymbol{u}$, t, etc. Then, a law for the involved interaction is also required (which can be gravitational, electromagnetic, etc.). The third law implies the conservation of the total momentum of an isolated system of interacting particles. In fact, the reciprocal forces $\boldsymbol{F}_{12}$ and $\boldsymbol{F}_{21}$ between two particles m_1

and m_2 satisfies $\boldsymbol{F}_{12}+\boldsymbol{F}_{21}=0$, since they are equal and opposite. If these are the only forces on each particle, we can use the second law to obtain $\mathrm{d}(\boldsymbol{p}_1+\boldsymbol{p}_2)/\mathrm{d}t=0$. Thus $\boldsymbol{p}_1+\boldsymbol{p}_2$ is a conserved quantity. This argument can be extended to prove the conservation of the total momentum of any isolated system of particles.

Classical mechanics allows for interacting forces at a distance. They are derived from potential energies depending on the distances between particles, which automatically provide interaction forces accomplishing Newton's third law.

1.2.2 Newton's Absolute Space

Newton's fundamental laws of dynamics are not formulated to be used in any frame. In fact, it is evident that the first law cannot be valid in any frame, since a constant velocity $\boldsymbol{u}$ in a frame S does not imply a constant velocity $\boldsymbol{u}'$ in another frame S'. This can be easily understood by considering cases where S' rotates or accelerates with respect to S. However, if S' translates uniformly with respect to S, either the particle has constant velocities $\boldsymbol{u}$, $\boldsymbol{u}'$ in both frames or in neither of them. Galilean addition of velocities (1.8) is a particular example of this general statement. In fact, Galilean transformations (1.4) were obtained for two equally oriented moving frames; thus, they are in relative translation (absence of relative rotation). Moreover, the translation is uniform, since the velocity $\boldsymbol{V}$ is constant. Thus $\boldsymbol{u}'$ is constant in (1.8) if and only if $\boldsymbol{u}$ is constant.

Although the principle of inertia cannot be valid in any frame, at least it is true that if it is valid in a frame S, then it will be valid in any other frame S' uniformly translating with respect to S. Can we extend this statement to the second law? The second law involves particle acceleration. In Galilean transformations, the acceleration is invariant. Besides, the forces in classical mechanics depend on distances (like gravitational and elastic forces) or relative velocities (like the viscous force on a particle moving in a fluid, which depends on the velocity of the particle relative to the fluid). Both the distances and the relative velocities are invariant under Galilean transformations. In this way, each side of the second law (1.11) is invariant under changes of frames in relative uniform translation. Therefore, the invariance of distances and time intervals, which leads to Galilean transformations, is a key piece in the Newtonian construction because it allows the second law to be valid in a family of frames in relative uniform translation. This is the family of *inertial frames*, and this is the content of the principle of relativity.

Principle of Relativity

The fundamental laws of physics have the same form in any inertial frame.

For instance, the same physical laws describe a free falling body both in a plane and at the Earth's surface. The principle of relativity in classical mechanics tells us that the state of motion of the frame cannot be revealed by a mechanical experiment: the result of the experiment will not depend on the motion of the frame because it is ruled by the same laws in all the inertial frames.

But how can we recognize whether a frame is inertial or not? We could effectively recognize a particle in rectilinear uniform motion; if we were sure that the particle is free of forces, then we would conclude that the frame is inertial. However, mechanics allows not only for contact forces but for forces *at a distance*. So how can we be sure that a particle is free of forces? Newton was aware of this annoying weakness of the formulation; he then considered that the laws of mechanics described the particle motion in the *absolute space*. Thus, the inertial frames are those fixed or uniformly translating with respect to Newton's absolute space.

While the inertial frames are defined by their states of motion with respect to Newton's absolute space, this (absolute) motion is not detectable, since the principle of relativity puts all the inertial frames on an equal footing; actually, only relative motions are detectable. Absolute space in classical mechanics plays the essential role of selecting the privileged family of inertial frames where the fundamental laws of physics are valid; but, surprisingly, it is not detectable. In some sense, absolute space *acts*, because it determines the inertial trajectories of particles, but it does not receive any reaction because it is immutable. *Leibniz* [1.5] criticized this feature of the Newtonian construction, by demanding that mechanics were aimed to describe relations among particles instead of particle motions in the absolute space. In practice, however, Newton's mechanics is successful because we can choose frames where the non-inertial effects are weak or can be understood in terms of *inertial forces* that result from referring the frame motion to another *more inertial* frame.

As advanced in Sect. 1.1.2, special relativity will abandon the invariance of distances and time intervals. Then, Galilean transformations will also be abandoned. This means that Newton's second law (1.11) and the character of fundamental forces will suffer a relativistic reformulation. However the inertial frames will still keep their privileged status devoid of a sound physical basis; this issue will be only re-elaborated in general relativity.

1.3 The Theory of Light and Absolute Motion

In the second half of the nineteenth century light was regarded as electromagnetic mechanical waves governed by Maxwell's laws. These waves were perturbations of a medium called ether; they propagate at the speed c relative to the ether. However, the ether could not be evidenced, neither directly nor indirectly. Several experiments did not succeed in revealing the Earth's motion relative to the ether (a sort of absolute motion), and some forced hypotheses about the interaction between matter and ether were introduced to give account of these null results.

1.3.1 The Finiteness of the Speed of Light

As mentioned in Sect. 1.1.2, velocities are not invariant in classical mechanics. Actually, only an infinite velocity would remain invariant under Galilean addition of velocities (1.8). Are there infinite speeds in nature? Many philosophers (Aristotle among them) thought that the speed of light was infinite. The issue of whether the speed of light was finite or infinite had been the object of debate since ancient times. In the seventeenth century, the question was still open. While Kepler and Descartes argued in favor of an infinite speed of light, Galileo proposed a terrestrial test that, however, was not suitable to determine such a large speed. However, at the end of the seventeenth century, contemporarily to Newton's development of mechanics, an answer came from the side of astronomy.

In 1676 *Rømer* [1.6] noted that the time elapsed between the observations of successive eclipses of Io – the innermost of Jupiter's great moons – was greater when the Earth traveled its solar orbit moving away from Jupiter and shorter when the Earth moved towards Jupiter. Rømer realized that such deviations in this otherwise periodical phenomenon were the sign of a finite speed of light. In fact, if the Earth were at rest, then we would observe one eclipse each 42.5 h (the orbital period of Io). However, if the Earth moves away from Jupiter, the time between two successive observations of the emersions of Io from the shadow cone will be enlarged; this happens because the light coming from the second emersion travels a longer distance at a finite velocity to reach the Earth. This delay, together with the length traveled by the Earth in 42.5 h, led to the first determination of the speed of light. By recording the accumulative delay of many successive eclipses, Rømer found that the light traveled the diameter of the Earth's orbit in 22 min (the actual value is 16 min) [1.7].

50 years later, *Bradley* [1.8] discovered the aberration of starlight. Bradley observed that the light coming from a star suffers annual changes of direction in the frame translating with the Earth. The nature of these changes highly disturbed Bradley because they unexpectedly differed from the stellar parallax he was looking for (a tiny effect only measured 100 years later). Eventually, Bradley concluded that the *stellar aberration* discovered by him was a consequence of the vector composition (1.8) between the speed of light and the Earth's motion around the Sun at $30\,\mathrm{km\,s^{-1}}$. By measuring the aberration angle, Bradley obtained the speed of light within an error of 1% [1.9]. In 1849 *Fizeau* [1.10] carried out the first terrestrial measurement of the speed of light. Like any finite velocity, the speed of light is not a Galilean invariant.

1.3.2 The Wave Equation

At the middle of the nineteenth century the dispute about the corpuscular or undulatory character of light seemed to be settled in favor of the wave theory of light. The corpuscular model sustained by Newton and many other scientists could not explain the totality of the luminous phenomena. In 1821 *Fresnel* [1.11] completed his wave theory of light, so giving a finished mathematical form to the undulatory model proposed by Huygens in 1678 [1.12]. This theory included the concepts of amplitude and phase to describe interference and diffraction; besides, light was presented as a transversal wave to explain the phenomena concerning polarization. In 1850 *Foucault* [1.13] measured the speed of light in water and verified the value c/n (n is the refractive index) as predicted by wave theory in opposition to the corpuscular model.

At that time, light waves were considered *matter* waves like sound or the waves on the water surface of a lake. Physics and mechanics were synonymous; so, any phenomenon was regarded as a mechanical phenomenon, and light did not escape the rule. Matter waves propagate in a material medium; they are but medium oscillations carrying energy. In the simplest cases, they are governed by the *wave equation*

$$\frac{1}{c_\mathrm{w}^2}\frac{\partial^2\psi}{\partial t^2} - \nabla^2\psi = 0\,, \tag{1.12}$$

where $\psi(t, \boldsymbol{r})$ represents the perturbation of the medium (for instance the longitudinal oscillations of density

and pressure when sound propagates in a gas, or the transversal displacement of a string in a musical instrument). Any function $\psi = \psi(x \pm c_\text{w} t)$ is a solution of the wave equation (1.12); it represents a perturbation that travels in the x-direction, without changing its form, at the constant speed $\pm c_\text{w}$. The general solution is a combination of solutions traveling in all directions.

The wave equation (1.12) was not written to be used in any inertial frame. It only describes the wave propagation in a frame fixed to the medium. In fact, the wave equation changes form under Galilean transformations. Let us take the x-sector of the Laplacian ∇^2 and write

$$\begin{aligned}\frac{1}{c_\text{w}^2}\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} &= \left(\frac{1}{c_\text{w}}\frac{\partial}{\partial t} - \frac{\partial}{\partial x}\right)\left(\frac{1}{c_\text{w}}\frac{\partial}{\partial t} + \frac{\partial}{\partial x}\right)\\ &= 4\frac{\partial}{\partial \xi}\frac{\partial}{\partial \eta}\,,\end{aligned} \tag{1.13}$$

where $\xi \equiv c_\text{w} t - x$, $\eta \equiv c_\text{w} t + x$ (or $c_\text{w} t = (\eta + \xi)/2$, $x = (\eta - \xi)/2$). This shows that the wave equation would keep its form in different inertial frames moving along the x-axis if $c_\text{w} t \pm x$ were proportional to $c_\text{w} t' \pm x'$; but this is not true in Galilean transformations (1.4), (1.6). The fact that the equation governing mechanical waves is fulfilled just in the frame where the medium is at rest does not imply the violation of the principle of relativity. The medium is a physical reason for privileging an inertial frame; furthermore, (1.12) will be accomplished whatever be the inertial frame where the medium is at rest. Actually, the wave equation for mechanical waves can be obtained from the fundamental laws of mechanics – which certainly accomplish the principle of relativity – under some assumptions valid in the frame fixed to the medium. In this derivation, the propagation velocity c_w results from the properties of the propagating media.

1.3.3 The Æther Theory

In Fresnel's theory, light was a mechanical wave that propagates in a medium called the *ether luminiferous*, and ψ was the *velocity of the ethereal molecules.* The speed of light c was a property of the ether. To be the seat of transversal waves, the ether had to be an elastic material; it was strange that no longitudinal waves existed in this elastic medium. Besides, to produce such enormous propagation velocity, the ether had to be extremely rigid. The ether had to fill the universe, because light propagates everywhere. It was logical to consider the ether as being at rest in Newton's absolute space; the ether became a sort of materialization of Newton's absolute space.

However, such an omnipresent substance should produce other mechanical effects, apart from the luminous phenomena. How can planets move through the ether without losing energy? Would the ether penetrate through the moving bodies without disturbing them or it would be dragged by them? If air is pumped out of a bottle, then the sound will cease to propagate inside the bottle; however, the light will still propagate, meaning that the ether was not evacuated together with the air (why?). The ether looked like an elusive intangible substance without any other effect than being the seat of the luminous phenomena.

1.3.4 Maxwell's Electromagnetism

In 1873 *Maxwell* [1.14] published his *Treatise on electricity and magnetism*, where electricity and magnetism appeared as two parts of a sole entity: the electromagnetic field. Maxwell's laws for the electromagnetic field contained as particular cases the well-known electrostatic interactions between charges and magnetostatic interactions between steady currents. However, Maxwell's very achievement was to discover that *variable* electric and magnetic fields – $\boldsymbol{E}$ and $\boldsymbol{B}$ – create each other. This mutual feedback between electricity and magnetism generates *electromagnetic waves.* In fact, in the absence of charges Maxwell's equations lead to wave equations (1.12), with the Cartesian components of $\boldsymbol{E}$ and $\boldsymbol{B}$ playing the role of ψ. In the electromagnetic wave equations the propagation velocity is $c = (\mu_\text{o}\varepsilon_\text{o})^{-1/2}$. In SI units, μ_o is chosen to define the unit of electric current, and ε_o is experimentally determined through electrostatic interactions; their values are $\mu_\text{o} = 4\pi \times 10^{-7}\,\text{N}\,\text{A}^{-2}$, $\varepsilon_\text{o} = 8.854187817 \times 10^{-12}\,\text{N}^{-1}\,\text{A}^2\,\text{m}^{-2}\text{s}^2$. To Maxwell's surprise, the value of c coincided with the already measured speed of light; so Maxwell concluded that light was an electromagnetic wave.

Maxwell conceived electromagnetic waves as a mechanical phenomenon in a propagating medium. Therefore, he believed that his equations were valid in a frame fixed to the medium. The recognition of light as an electromagnetic wave then identified the electromagnetic medium with the luminiferous ether. On the other hand, the action of the field on a charge q – the Lorentz force $\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{u} \times \boldsymbol{B})$ – depended on the velocity $\boldsymbol{u}$ of the charge. This velocity was regarded as the velocity of the charge with respect to the ether (the charge absolute velocity).

Differing from classical mechanics, Maxwell's electromagnetism will fit special relativity without changes. Einstein will defy the classical viewpoint by considering that Maxwell's equations should be valid in any inertial frame. If so, the speed of light would be invariant (i. e., it would have the same value in any inertial frame). To sustain this idea, Galilean transformations should be replaced with transformations leaving invariant the speed of light; this implies the abandonment of the classical invariance of distances and time intervals. In special relativity, Maxwell's electromagnetism will become a paradigmatic theory.

1.3.5 The Search for Absolute Motion

Although ether resisted a direct detection, at least it could be indirectly tested. In the second half of the nineteenth century, several experiments were aimed at testing the Earth's motion with respect to ether (the Earth's absolute motion). While c was considered the speed of light in the frame fixed to the ether, the speed of light in the Earth's frame should result from composing c with the Earth's absolute motion V, according to the Galilean addition of velocities (1.8). Therefore, some of these experiments were based on the time that the light takes to travel a round-trip along a straight path (the light comes back after being reflected by a mirror). To exemplify the idea, we will choose the path to be parallel to the (unknown) Earth's absolute motion. According to Galilean addition of velocities, the speed of light in the Earth's frame is $c - V$ when light leaves, and $c + V$ when light comes back. If l is the length the light covers in each journey, then the total time of the round-trip is

$$t = \frac{l}{c-V} + \frac{l}{c+V} = \frac{2l/c}{1-\frac{V^2}{c^2}} \,. \tag{1.14}$$

As can be seen, the Earth's absolute motion V enters the result as a correction of the second order in V/c. A correction of even order is expectable because the traveling time of a round trip (1.14) should not change if the Earth's motion were reversed. To be conclusive, the experiments should be able to detect at least a value $V/c \approx 10^{-4}$. This is because the Earth orbits the Sun at $30\,\text{km s}^{-1} \cong 10^{-4}c$; then, even if the Earth were at rest in the ether when the experiment is performed, it would move at $60\,\text{km s}^{-1}$ 6 months later. Therefore, any experimental array based on the traveling time (1.14) should reach a sensitivity of 10^{-8}. Such a strong constraint could be circumvented by experimental arrays sensitive to the change $V \to -V$; if so, the result could be of the first order in V/c. This the case of the experiment performed by *Hoek* [1.15] in 1868, where the symmetry $V \leftrightarrow -V$ is broken because one of the stretches of the round-trip was not in air but in water; in this stretch, the speed c/n replaces c in (1.14). However, Hoek's interferometric device was not effective for determining the Earth's absolute motion.

There were also two experiments, sensitive to the first order in V/c, that involved Snell's law. In 1871 *Airy* [1.16] measured Bradley's stellar aberration with a vertical telescope filled with water. Bradley had measured the annual variation of the aberration angle produced by the Earth's orbit around the Sun. This variation did not reveal the Earth's absolute motion $\boldsymbol{V}$ but just the changes of $\boldsymbol{V}$. Airy's experiment, instead, took into account that the aberration implied that the telescope was not oriented along the direction the light ray had in the ether's frame. If Snell's law were valid in the ether frame, then an additional refraction would take place when the light entered the water in the telescope. This additional refraction would change the view angle to the star by a quantity of the first order in V/c. Nevertheless, Airy's experiment did not reveal the Earth's absolute motion. Much earlier, in 1810, *Arago* [1.17] covered half of the objective of his telescope with a prism, to obtain a second image of the stars. To see the image through the prism, the telescope direction had to be corrected in an angle equal to the deviation angle of the prism. Arago believed that the light refraction in the prism could depend on the velocity of light relative to the prism, which results from the vector composition (1.8) of the speed of light with the absolute motion of the prism (i. e., the Earth's absolute motion). This effect could be revealed by observing stars in several directions to get different vector compositions. However, Arago did not notice any change in the deviation angle.

Fresnel [1.18] searched reasons for Arago's null result. In the context of the ether theory, he found that the null result could be explained, at the first order in V/c, by advancing a curious hypothesis: an (absolute) moving transparent substance partially drags the ether contained in its interior. The partial dragging is such that the phase velocity of light – the displacement per unit of time of the wave fronts –, as measured in the frame fixed to the universal ether (rather than the ether inside the substance) is not c/n but

$$u = \frac{c}{n} + (1 - n^{-2})\boldsymbol{V} \cdot \hat{\boldsymbol{n}} \,, \tag{1.15}$$

where $\hat{\boldsymbol{n}}$ is the propagation direction, $\boldsymbol{V}$ is the absolute motion of the transparent substance and n is its refrac-

tive index. In practice, Fresnel's dragging coefficient $f = 1 - n^{-2}$ caused the fulfillment of Snell's law in the frame fixed to the transparent substance (at the first order in V/c). Fresnel's hypothesis explained why Arago did not succeed in his endeavor: the deviation angle of the prism was always the one predicted by Snell's law, irrespective of the absolute motion of the prism. Besides, it also explained the null result in Airy's experiment because no additional refraction will be produced if Snell's law is valid in the frame fixed to the telescope (in this frame the ray of light and the telescope are equally oriented). Moreover, the partial dragging (1.15) cancels out the first order effects in the time (1.14) when one of the stretches is not in air but in another transparent substance; so, it also explained Hoek's null result (Hoek's device was not sensitive enough to test second order effects).

Fresnel's partial dragging of ether was measured by *Fizeau* [1.19] in 1851. Since special relativity will reject the existence of the ether, Fizeau's measurement will require a relativistic interpretation. On the other hand, the fulfillment of Snell's law in the frame fixed to the transparent substance is completely satisfactory in special relativity, because that is the only physically privileged frame. For a detailed analysis of the experiments pursuing the absolute motion in connection with Fresnel's hypothesis, see [1.20, 21].

1.3.6 Michelson–Morley Experiment

In 1881 Michelson designed an interferometer aimed to detect the Earth's absolute motion. In Michelson's interferometer the light traveled round-trips completely in air. So, the challenge was to achieve sensitivity of

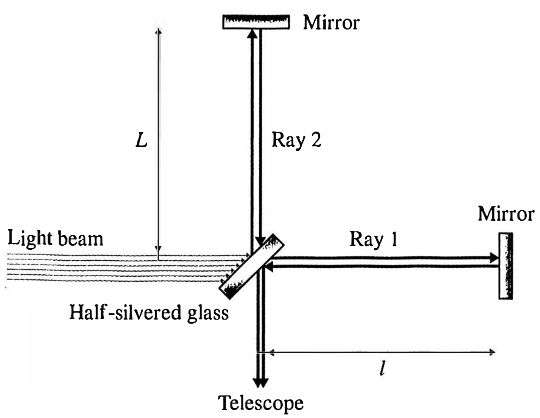


Fig. 1.2 Scheme of Michelson's interferometer

10^{-8}. Figure 1.2 shows the scheme of Michelson's interferometer. The beam of light emitted by an extensive source is split into two parts by a half-silvered glass plate. After traveling mutually perpendicular round-trips, both parts join again to be collected by a telescope where interference fringes are observed (Fizeau's fringes [1.22]). The fringes are caused by a slight misalignment of the mirrors; this implies that the images of the mirrors at the telescope form a wedge. The wedge causes that rays 1 and 2 arrive at the telescope with a phase-shift that changes according to the thickness of the wedge at the place where the rays bounced. So, the phase-shift will be different for each one of the rays in the beam; therefore, bright and dark fringes will be observed at the telescope. Notice that l and L do not need to be equal, but $2(l-L)$ should be smaller than the coherence length of light to preserve the interference pattern.

For each ray in the beam, the phase-shift between parts 1 and 2 determines whether they produce a bright or a dark fringe. This phase-shift results from the times t_1, t_2 the rays 1 and 2 employ to cover their respective round-trips; these times depend on the distances l, L and the velocities $\boldsymbol{u}'_1$, $\boldsymbol{u}'_2$ of the rays in the laboratory. $\boldsymbol{u}'_1$, $\boldsymbol{u}'_2$ are the result of the vector composition (1.8) between the speed c in the ether frame and the Earth's absolute motion $\boldsymbol{V}$; u'_1, u'_2 are clearly different, since the vector composition depends on the direction of each ray. Moreover, if the interferometer were gradually rotated then the velocities $\boldsymbol{u}'_1$, $\boldsymbol{u}'_2$ would gradually change. In this way, the rotation of the interferometer would affect the fringes: the position of the bright fringes would gradually displace. Instead, if the interferometer were at rest in the ether, then the fringes would not displace because rays 1 and 2 would travel at the speed c irrespective of the orientation of the interferometer. Thus, the displacement of the fringes would be the indication of the Earth's absolute motion.

Let us compute the times t_1, t_2 when the arm l is oriented along the still unknown absolute motion $\boldsymbol{V}$. In such a case, the ray 1 has speeds $c - V$, $c + V$, and the time t_1 is given by (1.14). On the other hand, the ray 2 is orthogonal to $\boldsymbol{V}$ in the laboratory frame; so the vector composition to obtain the value of $\boldsymbol{u}'_2$ is the one shown in Fig. 1.3. As can be seen, the ray 2 goes to the mirror and comes back with a speed $u'_2 = \sqrt{c^2 - V^2}$. Then, the round-trip along the arm L takes a time

$$t_2 = \frac{2L/c}{\sqrt{1 - \frac{V^2}{c^2}}} . \tag{1.16}$$

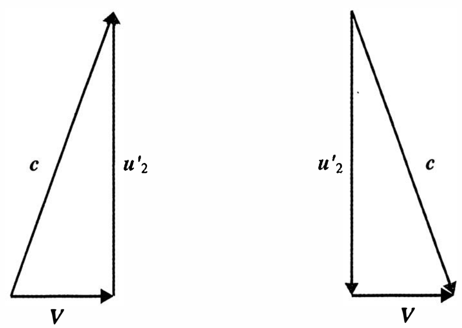


Fig. 1.3 Galilean composition of velocities for ray 2

The phase-shift is ruled by the time difference

$$\Delta t_{0^\circ} = t_1 - t_2 = \frac{2l/c}{1-\frac{V^2}{c^2}} - \frac{2L/c}{\sqrt{1-\frac{V^2}{c^2}}} \,. \tag{1.17}$$

If the interferometer is rotated 90°, then the arm L corresponding to ray 2, will be aligned with $\boldsymbol{V}$; so the result will be

$$\Delta t_{90^\circ} = t_1 - t_2 = \frac{2l/c}{\sqrt{1-\frac{V^2}{c^2}}} - \frac{2L/c}{1-\frac{V^2}{c^2}} \,. \tag{1.18}$$

Although the Earth's absolute motion $\boldsymbol{V}$ is unknown, a gradual rotation will make the interferometer pass through these two extreme values separated by a right angle. Thus a displacement of the fringes will be observed, in connection with the change of $t_1 - t_2$ given by

$$\Delta t_{90^\circ} - \Delta t_{0^\circ} = \frac{2}{c}(l+L)\left[\frac{1}{\sqrt{1-\frac{V^2}{c^2}}} - \frac{1}{1-\frac{V^2}{c^2}}\right] = -\frac{l+L}{c}\frac{V^2}{c^2} + O(V^4c^{-4}) \,. \tag{1.19}$$

This change is equivalent to the displacement of $N = c|\Delta t_{90^\circ} - \Delta t_{0^\circ}|/\lambda = (l+L)/\lambda \times V^2/c^2$ fringes (λ is the light wavelength).

After a failed attempt in 1881, Michelson joined Morley to improve the experimental sensitivity. In 1887 they possessed an interferometer whose arms were 11 m long (this was achieved by means of multiple reflections in a set of mirrors). Then, at least a result of $N \cong 0.4$ was expected. However, no displacement of fringes was observed [1.23–25]. Michelson was convinced that the null result meant that the Earth carried a layer of ether stuck to its surface. If so, the experiment would have been performed at rest in the local ether, which would explain the null result. *Lodge* [1.26] tried to confirm this hypothesis by unsuccessfully looking for effects due to the ether stuck to a fast rotating wheel. In a revival of the corpuscular model, *Ritz* [1.27] then proposed that light propagates with speed c relative to the source. This hypothesis combined with other assumptions about the behavior of light when reflected by a mirror (*emission theories*) would explain the null result of Michelson–Morley's experiment with a source at rest in the laboratory, but is refuted by a varied body of experimental evidence [1.28–30].

1.3.7 FitzGerald–Lorentz Length Contraction

Lorentz thought that Michelson–Morley's null result could be understood in a very different way. He considered that a body moving in the ether suffered a length contraction due to its interaction with the ether. The interaction would contract the body along the direction of its absolute motion $\boldsymbol{V}$, but the transversal dimensions would not undergo any change. In fact, if the contraction factor $\sqrt{1-V^2c^{-2}}$ is applied to l in (1.17) and L in (1.18) (i. e., the dimensions along the absolute motion direction in each case), then both time differences will result to be equal, and the expression (1.19) will vanish. Lorentz's proposal of 1892 [1.31] had been independently advanced by *FitzGerald* [1.32] 3 years before. This proposal did not mean the abandonment of the belief in the invariance of lengths. The contraction was a dynamical effect; it depended on an objective phenomena: the interaction between two material substances. The contraction had to be observed in any frame, and all the frames had to agree about the value of the contracted length.

The idea that light was a material wave (i. e., the idea that Maxwell's laws were written to be used only in the ether frame) and the belief in the invariance of distances and time intervals led physics to a blind alley. While complicated dynamical explanations were elaborated to interpret experimental results, like Fresnel's partial dragging of ether and FitzGerald–Lorentz length contraction caused by the ether, the experimental results were not so complicated; they just said that the absolute motion cannot be detected. However, unless physics were to get rid of some classical misconceptions, such a reasonable conclusion would not fit with its theoretical body.

1.4 Einstein's Special Relativity

In 1905 *Einstein* postulated that [1.33]

> *the same laws of electrodynamics and optics will be valid for all frames of reference for which the equations of mechanics hold good.*

In this way, Einstein proclaimed that Maxwell's electromagnetism does not possess a privileged system; Maxwell's laws can be used in any inertial frame. Thus, Einstein raised Maxwell's laws to the status of fundamental laws satisfying the principle of relativity (as stated in Sect. 1.2.2). In doing so, Einstein closed the possibility of detecting the state of motion of an inertial frame by electromagnetic means. The ether does not exist; the electromagnetic waves are not material waves. The inertial frames are not endowed with a property V (its absolute motion or the *ether wind*); only the velocity describing the relative motion between inertial frames makes physical sense. Besides, the Snell's law is valid in the frame where the refracting substance is at rest, whatever this frame is.

An immediate consequence of the use of Maxwell's laws in any inertial frame is that light in vacuum propagates at the speed c in any inertial frame; c is an invariant velocity (*light is always propagated in empty space with a definite velocity c which is independent of the state of motion of the emitting body* [1.33]). The existence of an invariant velocity implies that Galilean addition of velocities is a classical misconception to be got rid of; such a step entails the revision of the classical belief in the invariance of distances and time intervals.

1.4.1 Relativistic Length Contractions and Time Dilations

We will re-elaborate the transformations of spacetime coordinates without prejudging about the behavior of distances and time intervals, but subordinating them to the invariance of the speed of light. Figure 1.4 shows a particle traveling between the ends of a bar, as seen in the frame where the bar is fixed and the frame where the particle is fixed. The relative motion bar-particle is characterized by the sole velocity V. It is useful to call *proper length* L_o the length of the bar at rest. Notice that, since all inertial frames are on an equal footing, the length of the bar will be L_o in any inertial frame where the bar is at rest. Instead, we could expect a different length $L(V)$ in a frame where the bar moves lengthways at a relative velocity V. For this reason, in Fig. 1.4 the bar is represented with different lengths in each frame.

In the frame fixed to the bar (*proper frame* of the bar) the particle takes a time Δt to cover the length L_o; then, it is $V = L_o/\Delta t$. On the other hand, in the frame fixed to the particle, the ends of the bar take a time $\Delta\tau$ to pass in front of the particle; then $V = L/\Delta\tau$. We should not prejudge the nature of time; then, we are opening the possibility that the time interval between the same pair of events be different in each frame. It is also useful to call *proper time* $\Delta\tau$ the time between events as measured in the frame where the events occur at the same place (if such a frame exists). In our case, the events are the passing of each end of the bar in front of the particle; they occur at the same place in the frame where the particle is fixed. So, we have computed the same value of V with lengths and times measured in two frames that relatively moves at a velocity V. Thus, we conclude that

$$\frac{L_o}{L} = \frac{\Delta t}{\Delta\tau} . \tag{1.20}$$

Each side of (1.20) can only depend on the relative velocity between the considered frames. Then, (1.20) says that each side is the same function of V

$$\frac{L_o}{L} = \gamma(V) , \quad \frac{\Delta t}{\Delta\tau} = \gamma(V) . \tag{1.21}$$

In classical physics $\gamma(V)$ is assumed to be 1. On the contrary, in special relativity the value of $\gamma(V)$ will be subordinated to the invariance of the speed of light. It should be remarked that (1.21) is not deprived of assumptions about the nature of spacetime. In fact, the quotients L_o/L and $\Delta t/\Delta\tau$ can also depend on the event of the spacetime where the measurements take place and the orientation of the bar. Equation (1.21) actually assumes that spacetime is homogeneous and isotropic; these assumptions will be revised in general relativity.

On the one hand, (1.21) expresses the relation between the length L of a bar moving at a velocity V and its proper length L_o. On the other hand, (1.21) expresses the relation between the times elapsed between two

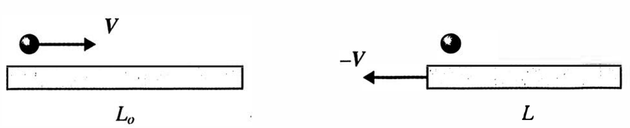


Fig. 1.4 Relative motion bar-particle in the proper frames of the bar (*left*) and the particle (*right*)

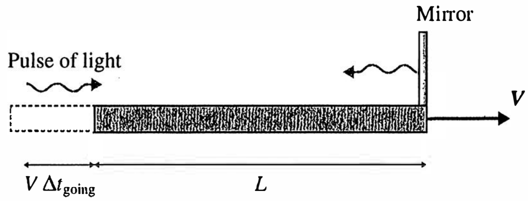


Fig. 1.5 A light pulse traveling a round trip between the ends of a bar, as regarded in the frame where the bar moves with velocity V

events as measured in the frame where they occur at the same place (proper time $\Delta\tau$) and another frame moving at a velocity V relative to the former one (Δt). As (1.21) shows, both ratios are strongly interconnected.

The relations (1.21) are independent of the particular case examined in Fig. 1.4. To obtain $\gamma(V)$ we will now study a case involving the speed of light, where the relations (1.21) will also enter into play. Figure 1.5 shows a bar of proper length L_o supporting a source of light and a mirror at its ends. Let us consider the time elapsed between the emission of a pulse of light from the source and its return to the source. Both events occur at the same place in the proper frame of the bar; then, the proper time $\Delta\tau$ is the time the light takes to cover the distance $2L_o$ at the speed c

$$c\Delta\tau = 2L_o\,. \tag{1.22}$$

In another frame where the bar moves at a velocity V (but light still propagates at the speed c), we will decompose the time between events as $\Delta t = \Delta t_{\text{going}} + \Delta t_{\text{returning}}$. When light goes towards the mirror at the speed c it covers the distance L plus the displacement of the mirror $V\Delta t_{\text{going}}$. Instead, when light returns to the source it covers the distance $L - V\Delta t_{\text{returning}}$ due to the displacement of the source. Therefore,

$$\begin{aligned} c\Delta t_{\text{going}} &= L + V\Delta t_{\text{going}}\,, \\ c\Delta t_{\text{returning}} &= L - V\Delta t_{\text{returning}}\,. \end{aligned} \tag{1.23}$$

Solving these equations for $c\Delta t_{\text{going}}$, $c\Delta t_{\text{returning}}$ one obtains

$$\begin{aligned} c\Delta t &= c\Delta t_{\text{going}} + c\Delta t_{\text{returning}} = \frac{cL}{c-V} + \frac{cL}{c+V} \\ &= \frac{2L}{1-\frac{V^2}{c^2}}\,. \end{aligned} \tag{1.24}$$

We divide (1.22) and (1.24) and use (1.21) to obtain the function $\gamma(V)$

$$\gamma(V) = \frac{1}{\sqrt{1-\frac{V^2}{c^2}}}\,. \tag{1.25}$$

Then, replacing in (1.21) we obtain the expressions for the relativistic *length contraction* and *time dilation*

$$\begin{aligned} L(V) &= L_o\sqrt{1-\frac{V^2}{c^2}}\,, \\ \Delta t_V &= \frac{\Delta\tau}{\sqrt{1-\frac{V^2}{c^2}}}\,. \end{aligned} \tag{1.26}$$

Noticeably, the relativistic length contraction has the same form as that proposed by FitzGerald and Lorentz to explain the null result of the Michelson–Morley experiment. However, its meaning is completely different. Lorentz considered that the contraction was a dynamical effect produced by the interaction between a body and the ether. For Lorentz, V in (1.26) was the velocity of the body with respect to the ether, and the contraction was measured in all the frames. In relativity, instead, the length contraction is a kinematical effect. The bar looks contracted whatever the frame be where it moves at the velocity V; moreover, it has its proper length L_o whatever the frame be where the bar is at rest.

Length contractions and time dilations are not perceptible in our daily life because we compare frames moving at relative velocities $V \ll c$. One of the first direct evidences of this phenomenon came from measuring the length traveled by decaying particles moving at a speed close to c, as compared to their half-life measured at rest [1.34].

1.4.2 Lengths Transversal to the Relative Motion

The device of Fig. 1.5 is also useful to explore the behavior of the dimensions transversal to the relative motion. Figure 1.6 shows the device put in a direction orthogonal to the relative motion. Equation (1.22) is still valid in the proper frame of the bar. In a frame where the bar transversally displaces at the velocity V, the ray of light will travel along an oblique direction (this is nothing but the aberration due to the composition of motions). When the pulse of light goes towards the mirror, it covers in a time Δt_{going} the hypotenuse

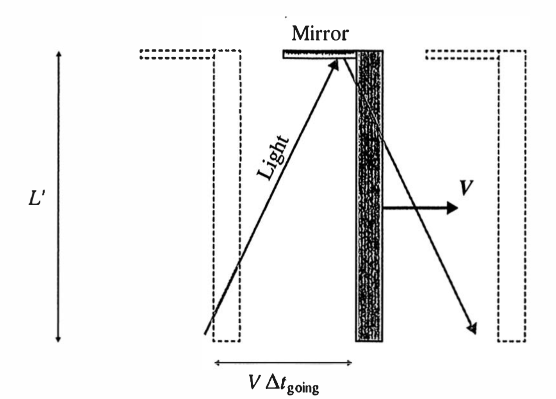


Fig. 1.6 Round trip of light between the ends of a bar, as regarded in a frame where the bar displaces transversally at velocity V

of a right triangle whose legs are $V\Delta t_{going}$ and L'. Since the light travels at the speed c in any frame, we obtain

$$(c\Delta t_{going})^2 = L'^2 + (V\Delta t_{going})^2 \,. \tag{1.27}$$

Notice the use of Pythagoras' theorem in this expression. This means that we assume the space is endowed with a flat geometry; this assumption will be revised in general relativity. Due to the symmetry of the path traveled by the light, it is $\Delta t = 2\Delta t_{going}$, and then

$$c\Delta t = \frac{2L'}{\sqrt{1-\frac{V^2}{c^2}}} = \gamma(V) 2L' \,. \tag{1.28}$$

We divide (1.22) and (1.28) and use (1.21) to obtain that transversal lengths are invariant

$$L' = L_o \,. \tag{1.29}$$

1.4.3 Lorentz Transformations

We are now in a position to reanalyze the transformation of the Cartesian coordinates of an event. Let us come back to (1.3) where the relation between $d_{O'P}|_S$ and x' is pending. By definition, the coordinate x' is the distance measured by a rule fixed to the frame S': $x' = d_{O'P}|_{S'}$. This rule looks contracted in the frame S; according to (1.26) it is $d_{O'P}|_S = \sqrt{1-V^2c^{-2}}x'$. Therefore,

$$x' = \gamma\,(x - Vt) \tag{1.30}$$

is the transformation that replaces (1.4a). We can now reproduce the argument of Sect. 1.1.3 to obtain the transformation of the time coordinate of an event. Since frames S and S' are on an equal footing, the inverse transformations have the same form, except for the change $V \to -V$. In particular, the inverse transformation of (1.30) is

$$x = \gamma\,(x' + Vt') \,. \tag{1.31}$$

Equation (1.30) can be replaced in (1.31) to solve t' as a function of t, x. Moreover, due to the relativistic invariance of the transversal lengths (1.29), the transformations (1.4b), (1.4c) remain valid. Finally, we obtain the *Lorentz transformations*

$$ct' = \gamma\,(ct - \beta x) \,, \tag{1.32a}$$

$$x' = \gamma\,(x - \beta ct) \,, \tag{1.32b}$$

$$y' = y \,, \tag{1.32c}$$

$$z' = z \,, \tag{1.32d}$$

where $\beta \equiv V/c$, $\gamma = \left(1-\beta^2\right)^{-1/2}$. Lorentz transformations (1.32) express the relativistic transformation of the coordinates of an event, when the inertial frame S is changed for an equally oriented inertial frame S' that moves along the (shared) x-axis at the relative velocity V. Notice that, since the transformation (1.32) is homogeneous, the same event is the coordinate origin for S and S'. Figure 1.7 shows the lines $t' = constant$ (i. e., $ct = \beta x + const$) and $x' = constant$ (i. e., $ct = x/\beta + const$) in the plane ct versus x. Figure 1.7 also

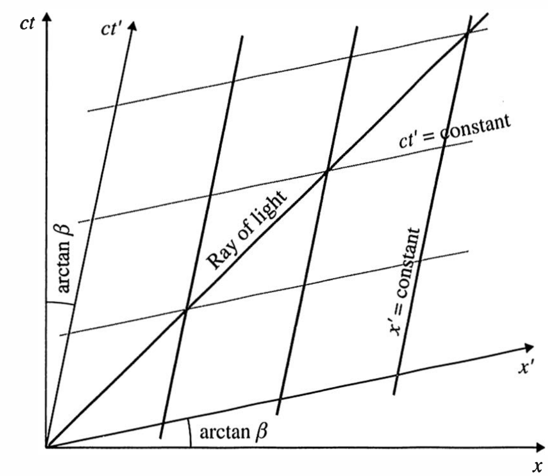


Fig. 1.7 Coordinate lines of S' in the plane ct versus x

displays a ray of light passing the coordinate origin and traveling in the x-direction; its world-line is a straight line at 45° because $\Delta x = c\Delta t$. Galilean transformations (1.4) are the limit $c \to \infty$ of Lorentz transformations (1.32).

The transformations (1.32) were independently obtained by *Lorentz* [1.35, 36] and *Larmor* [1.37] as the linear coordinate changes leaving the form of Maxwell's wave equations invariant (see also *Voigt* [1.38]). In fact, the *null coordinates* $\xi \equiv ct - x$, $\eta \equiv ct + x$ transform as $\xi' = \gamma(1+\beta)\xi$, $\eta' = \gamma(1-\beta)\eta$, thus leaving invariant the form of the wave equation (1.13) for $c_w = c$. In other words, the d' Alembertian operator

$$\Box \equiv \frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2 \tag{1.33}$$

is invariant under transformations (1.32). In 1905 *Poincaré* [1.39] underlined the group properties of relations (1.32) and called them Lorentz transformations.

In 1905 Einstein re-derived the Lorentz transformations and gave to t' the rank of real time measured by clocks at rest in S'. In Einstein's special relativity the physical equivalence of the inertial frames, which is the content of the principle of relativity, means that the fundamental laws of physics keep their form under Lorentz transformations rather than Galilean transformations. Maxwell's laws accomplish this relativistic version of the principle of relativity, once the transformations of the fields are properly defined. Actually, Maxwell's electromagnetism is the paradigm of a relativistic theory. The electromagnetic Lorentz force is a typical relativistic force; its magnetic part depends on the charge velocity relative to the inertial frame. However, which part of the field is electric and which one is magnetic depends on the frame as well; even if the force is entirely electric in a given frame, it will have a magnetic part in another frame. On the contrary, classical mechanics fulfilled the principle of relativity under Galilean transformations; then, mechanics needed a reformulation to accommodate to the relativistic meaning of the principle of relativity.

1.4.4 Relativistic Composition of Motions

The composition of motions that replaces the Galilean addition of velocities is obtained by differentiating (1.32) and taking quotients. Notice that

$$\mathrm{d}t' = \gamma(\mathrm{d}t - \beta c^{-1}\,\mathrm{d}x) = \gamma\left(1 - \beta c^{-1} u_x\right)\mathrm{d}t\,. \tag{1.34}$$

Therefore,

$$u'_x = \frac{\mathrm{d}x'}{\mathrm{d}t'} = \gamma\left(\frac{\mathrm{d}x}{\mathrm{d}t'} - V\frac{\mathrm{d}t}{\mathrm{d}t'}\right) = \frac{u_x - V}{1 - \beta c^{-1} u_x}\,, \tag{1.35a}$$

$$u'_y = \frac{\mathrm{d}y'}{\mathrm{d}t'} = \frac{\sqrt{1-\beta^2}\,u_y}{1 - \beta c^{-1} u_x}\,, \quad u'_z = \frac{\mathrm{d}z'}{\mathrm{d}t'} = \frac{\sqrt{1-\beta^2}\,u_z}{1 - \beta c^{-1} u_x}\,. \tag{1.35b}$$

The procedure can be repeated to transform the accelerations. Contrasting with Galilean transformations, the acceleration is far from being invariant under Lorentz transformations.

Equations (1.35a) and (1.35b) can be combined to obtain $u'^2 = u'^2_x + u'^2_y + u'^2_z$; it is easy to verify that

$$1 - \frac{u'^2}{c^2} = \frac{1-\beta^2}{(1-\beta c^{-1}u_x)^2}\left(1 - \frac{u^2}{c^2}\right). \tag{1.36}$$

Since $\beta < 1$ (otherwise Lorentz transformations would be ill-defined), both hand sides of (1.36) have the same sign. Therefore u and u' are both lower, equal to, or bigger than c; this is an invariant property of speed.

As an application of transformations (1.35a) and (1.35b), let us compute the speed of light when light propagates in a transparent substance that moves at the velocity V; then, $u'_x = c/n$, where n is the refractive index. We will use the inverse transformations to obtain u_x (i. e., we change V for $-V$ in (1.35a))

$$u_x = \frac{c}{n}\frac{1+\frac{nV}{c}}{1+\frac{V}{nc}} \approx \frac{c}{n}\left(1+\frac{nV}{c}\right)\left(1-\frac{V}{nc}\right) \approx \frac{c}{n} + (1-n^{-2})V\,. \tag{1.37}$$

This result has the same form as Fresnel's partial dragging. However, V in (1.37) is not the velocity of the transparent substance with respect to the ether; it is the motion of the transparent substance relative to an arbitrary inertial frame. What Fizeau measured in 1851 was a relativistic composition of motions.

1.4.5 Relativity of Simultaneity. Causality

Two events 1 and 2 (two points in the spacetime) are simultaneous if they have the same time coordinate:

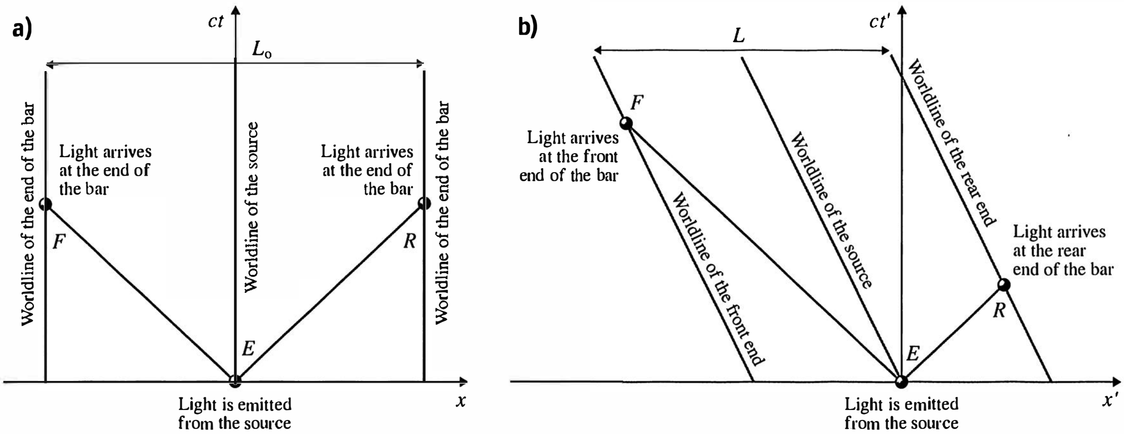


Fig. 1.8 (**a**) In the proper frame of the bar, the pulses of light arrive at the ends of the bar at the same time. (**b**) In a frame where the bar is moving, the light arrives at the rear end before than the front end. In both frames the speed of light is c (the rays of light are lines at 45°)

$t_1 = t_2$. In classical physics time is invariant, so the simultaneity of events possesses an absolute meaning. However, in special relativity $t_1 = t_2$ does not imply $t'_1 = t'_2$. Then the simultaneity acquires a relative meaning; it is frame-dependent. In fact, the pairs of events that are simultaneous in the frame S lie on horizontal lines ($t = constant$) in Fig. 1.7; these lines cross the $t' = constant$ lines. Therefore, simultaneous events in S have different time coordinates t' in S'.

To understand why simultaneity is relative in special relativity, let us consider a bar of proper length L_o that is equipped with a source of light at its center. In the proper frame of the bar, a pulse of light will arrive simultaneously at both ends of the bar, because it covers the same distance $L_o/2$ at the same speed c in both directions. In another frame the bar is moving but light still propagates at the speed c in any direction. Thus, the pulse will arrive first at the rear end of the bar because this end moves towards the pulse of light. Then, the same pair of events (the arrival of the light to the ends of the bar) is not simultaneous in a frame where the bar is moving. Moreover, since which end is at rear depends on the direction of the motion (i. e., it depends on the frame), the temporal order of this kind of events can be inverted by changing the frame.

Figure 1.8 shows the world-lines of the ends of the bar and the pulses of light both in the bar proper frame S and a frame S' where the bar moves to the left (then S' moves to the right relative to S, so it is $\beta > 0$). In Fig. 1.8a the ends of the bar are described by vertical world-lines because the positions x are fixed. In Fig. 1.8b the world-lines have a slope corresponding to the velocity $-V$ that the bar has in the frame S'. In both frames the light travels at the speed c. Events R and F are simultaneous in the proper frame of the bar (Fig. 1.8a) and they occur at a distance L_o. Then, $\Delta t = 0$, $\Delta x = -L_o$ ($\Delta t = t_F - t_R$, etc.). The time elapsed between R and F in the frame S' can be obtained by means of Lorentz transformations. Since Lorentz transformations are linear, they are equally valid for the differences of coordinates of a pair of events. So, (1.32a) also means

$$c\Delta t' = \gamma(c\Delta t - \beta\Delta x) . \qquad (1.38)$$

Then it is $c\Delta t' = \gamma\beta L_o$ in Fig. 1.8b. This result can be also achieved by applying elementary kinematics in the frame S' and using the length contraction $L = \gamma^{-1}L_o$.

In any case, (1.38) says that Δt and $\Delta t'$ cannot both be zero (apart from the case where the events are coincident). Moreover, Δt and $\Delta t'$ in (1.38) can even have opposite signs, which would amount to the inversion of the temporal order of events. This alteration of the temporal order in Lorentz transformations would be acceptable only for pairs of events without causal relation; otherwise it would constitute a violation of causality. Remarkably, the violation of causality is prevented because the speed of light cannot be exceeded in special relativity. As it will be shown in Sect. 1.5, c is an unreachable limit velocity for massive particles. Con-

sistently, it is $V/c = \beta < 1$ in Lorentz transformations. Therefore, those pairs of events such that $|\Delta \boldsymbol{r}| > c|\Delta t|$ cannot be in causal relation because neither particles nor rays of light can connect them. For instance, in Fig. 1.8 the events R and F cannot be in causal relation because their spatial separation is larger than their temporal separation. This property does not depend on the chosen frame, as can be checked in the transformations (1.32) or inferred from (1.36). On the contrary, pairs of events having $|\Delta \boldsymbol{r}| \leq c|\Delta t|$ can be causally connected. However, in this case, it results that $|\beta \Delta x| < |\Delta| \leq c\Delta t$. Thus $|\beta \Delta x|$ is not large enough to invert the temporal order in (1.38); so causality is preserved.

The relativity of simultaneity is usually the explanation to some *paradoxes* in special relativity. For instance, let us consider two bars having the same length if compared at relative rest. Then, if they are in relative motion, each one will appear shorter when regarded from the proper frame of the other one. How can this make sense? It makes sense because the length of a bar results from comparing the simultaneous positions of its ends. Since the simultaneity is not absolute in special relativity, a length measurement performed in S is not consistent in S'.

1.4.6 Proper Time of a Particle

While those events having $|\Delta \boldsymbol{r}| > c|\Delta t|$ admit a frame where they occur at the same time (or, moreover, frames where their temporal order is inverted), those events having $|\Delta \boldsymbol{r}| < c\Delta t$ admit a frame where they occur at the same place. This is a consequence of the symmetric form of (1.32a) and (1.32b). From a more physical standpoint, the events having $|\Delta \boldsymbol{r}| < c\Delta t$ can be joined by a uniformly moving particle. The proper frame of the particle effectively realizes the inertial frame where both events occur at the same place: the events occur at the (fixed) position of the particle. These observations show that the concept of proper time, as defined in Sect. 1.4.1, applies to pairs of events whose spatial separation is smaller than the temporal separation.

In general, any moving particle causally connects events. Figure 1.9 shows the world-line of a particle that moves nonuniformly. Since the world-line cannot exceed the angle of 45° characterizing the speed of light, any pair of events on the world-line of the particle will satisfy $|\Delta \boldsymbol{r}| \leq c\Delta t$. Let us consider two infinitesimally closed events, like those shown in Fig. 1.9 corresponding to the times t and $t + \mathrm{d}t$. The frame where these two events occur at the same place is the proper frame of the

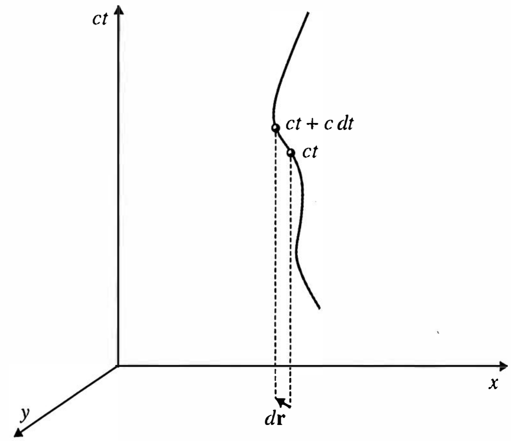


Fig. 1.9 Two infinitesimally closed events belonging to the world-line of a nonuniformly moving particle. They are causally connected: $|\mathrm{d}\boldsymbol{r}| < c\mathrm{d}t$

particle moving at the speed $u(t)$. Let us rewrite (1.36) with the help of (1.34) to obtain

$$\sqrt{1 - \frac{u'^2}{c^2}}\,\mathrm{d}t' = \sqrt{1 - \frac{u^2}{c^2}}\,\mathrm{d}t\,. \tag{1.39}$$

As can be seen, this is a combination of speed and time of travel, which has the same value in any frame: it is invariant. By comparing this with (1.26) one realizes that the invariant (1.39) is nothing but the proper time elapsed between the infinitesimally closed events. In other words, (1.39) is the time measured by a clock fixed to the particle; it is the proper time of the particle

$$\mathrm{d}\tau = \sqrt{1 - \frac{u^2}{c^2}}\,\mathrm{d}t = \gamma(u)^{-1}\,\mathrm{d}t\,. \tag{1.40}$$

This expression can be integrated along the world-line to obtain the total time measured by a clock that moves between a given pair of causally-connectable events. Clearly, the integral depends on the world-line the clock uses to join the initial and final events (it depends on the function $u(t)$). It is easy to prove that the total proper time is maximized along an inertial world-line. This result is related to the so-called *twin paradox*. The paradox refers to twin brothers who separate because one of them has a space voyage. When they meet again, the *inertial* brother who remained at the Earth is older than the astronaut. Actually this result is not paradox-

ical; the brothers are not on an equal footing because special relativity confers a privileged status to inertial frames.

1.4.7 Transformations of Rays of Light

Let us consider a monochromatic plane solution of (1.12) for waves traveling at the speed of light

$$\psi \propto \exp\left[\mathrm{i}\frac{2\pi\nu}{c}(ct - \hat{\boldsymbol{n}}\cdot\boldsymbol{r})\right], \tag{1.41}$$

where the unitary vector $\hat{\boldsymbol{n}}$ is the propagation direction, and ν is the frequency. Let us use the inverse Lorentz transformations to rewrite the phase of the wave in terms of coordinates in S'

$$\begin{aligned}\nu(ct - \hat{\boldsymbol{n}}\cdot\boldsymbol{r}) &= \nu\,\big(\gamma(ct' + \beta x') \\ &\quad - n_x\gamma(x' + \beta ct') \\ &\quad - n_y y' - n_z z'\big) \\ &= \gamma(1 - \hat{\boldsymbol{n}}\cdot\boldsymbol{V}c^{-1})\nu ct' \\ &\quad - \nu\,\big(\gamma(n_x - \beta)x' + n_y y' \\ &\quad + n_z z'\big)\,. \end{aligned} \tag{1.42}$$

Since the d'Alembertian operator (1.33) keeps the same form if rewritten in coordinates of S', the result (1.42) should be reinterpreted as $\nu'(ct' - \hat{\boldsymbol{n}}'\cdot\boldsymbol{r}')$. Therefore, one obtains the relativistic Doppler effect and light aberration.

Doppler Effect for Light

The frequency in the frame S' is

$$\nu' = \gamma(1 - \hat{\boldsymbol{n}}\cdot\boldsymbol{V}c^{-1})\nu\,. \tag{1.43}$$

Factor γ is absent in the classical Doppler effect. It implies that the frequency shift exists even if the propagation direction is orthogonal to $\boldsymbol{V}$ (transversal Doppler effect) due to time dilation. The first verification of the relativistic Doppler frequency shift was made in 1938 [1.40].

Light Aberration

Moreover, it is $n'_x = (\nu/\nu')\gamma(n_x - \beta) = (n_x - \beta)/(1 - \beta n_x)$. If θ is the angle between the ray of light and the x-axis, then it is $n_x = \cos\theta$. Thus, the propagation direction transforms as

$$\cos\theta' = \frac{\cos\theta - \beta}{1 - \beta\cos\theta}\,. \tag{1.44}$$

The aberration angle is $\alpha \equiv \theta' - \theta$; α is very small whenever it is $\beta \ll 1$. So, we can approach $\cos\theta' = \cos(\theta + \alpha) \approx \cos\theta - \alpha\sin\theta$. Moreover, the right-hand side of (1.44) can be approached by $\cos\theta - \beta\sin^2\theta$. Therefore,

$$\alpha \approx \beta\sin\theta\,, \tag{1.45}$$

which is the Galilean approach Bradley used to obtain the speed of light from the annual variation of the starlight aberration.

1.5 Relativistic Mechanics

While the principle of inertia remains valid in special relativity, Newton's second law has to be reformulated because it does not satisfy the principle of relativity under Lorentz transformations (forces behave differently than accelerations under Lorentz transformations). Relativistic mechanics can be constructed from a Lorentz-invariant variational principle whose functional action reproduces Newtonian behavior at low velocities. In special relativity, energy and momentum are strongly related. Momentum is conserved in any frame if and only if the energy is also conserved. When particles collide, the conservation of the relativistic energy takes the role of classical mass conservation. However, the relativistic energy is a combination of mass and kinetic energy; so, mass can be converted in kinetic energy (or other energies, like the electromagnetic energy associated with photons) and vice versa. Classical interactions at a distance are excluded because the relativity of simultaneity prevents nonlocal conservations of energy–momentum. Instead, the interactions *at a distance* are realized through mediating fields carrying energy–momentum that locally interact with the particles.

1.5.1 Momentum and Energy of a Particle

Variational principles are an outstanding tool to build dynamical theories in Physics. They rest on the stationarity of a functional action. The resulting Lagrange dynamical equations will fulfill the principle of relativ-

ity under Lorentz transformations whenever the action is Lorentz-invariant. This feature guarantees that different inertial frames will agree about the stationarity of the action. Thus, the same set of equations of motion will be valid in all the inertial frames.

Let us start by building the action of a free particle. This action not only has to be Lorentz invariant but must be equivalent to the classical action when $|\boldsymbol{u}| \ll c$. The (invariant) proper time along the particle world-line (1.40) is the right choice for the functional action of the free particle

$$\begin{aligned} S_{\text{free}}[\boldsymbol{r}(t)] &= -mc^2 \int \mathrm{d}\tau \\ &= -mc^2 \int \sqrt{1-\frac{|\boldsymbol{u}|^2}{c^2}}\,\mathrm{d}t \\ &= -mc^2 \int \gamma(u)^{-1}\,\mathrm{d}t\,. \end{aligned} \tag{1.46}$$

When $|\boldsymbol{u}| \ll c$ the Lagrangian $L = -mc^2(1-u^2/c^2)^{1/2}$ goes to $L \approx -mc^2 + (1/2)mu^2$. By differentiating the Lagrangian L with respect to $\boldsymbol{u}$ one obtains the conjugate momentum $m\gamma(u)\boldsymbol{u}$ of a free particle. One then defines the *momentum* of the particle as

$$\boldsymbol{p} \equiv m\gamma(u)\boldsymbol{u} = m\gamma(u)\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t} = m\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}\tau} \tag{1.47}$$

(the last step results from (1.40)), which goes to the classical momentum $m\boldsymbol{u}$ when $|\boldsymbol{u}| \ll c$.

Since $\mathrm{d}\tau$ is invariant (1.39), the change of $\boldsymbol{p}$ under Lorentz transformations emanates from the behavior of $\mathrm{d}\boldsymbol{r}$. A Lorentz transformation mixes $\mathrm{d}\boldsymbol{r}$ with $c\,\mathrm{d}t$. Then $\boldsymbol{p}$ will be mixed with $mc\,\mathrm{d}t/\mathrm{d}\tau$, which is a quantity intimately related to the energy. In fact, the Hamiltonian of the free particle is

$$\begin{aligned} H &= \boldsymbol{u}\cdot\boldsymbol{p} - L = m\gamma(u)u^2 + mc^2\gamma(u)^{-1} \\ &= mc^2\gamma(u)\left(\frac{u^2}{c^2} + \gamma(u)^{-2}\right) = mc^2\gamma(u)\,. \end{aligned} \tag{1.48}$$

Then, we define the *energy* of the particle as

$$E \equiv m\gamma(u)c^2\,. \tag{1.49}$$

The energy E is a combination of *energy at rest* mc^2 and kinetic energy. In fact, by Taylor expanding (1.49) we obtain

$$E = mc^2 + \frac{1}{2}mu^2 + \cdots \equiv mc^2 + T\,, \tag{1.50}$$

where T is the kinetic energy of the particle in special relativity (at low velocities, it coincides with the classical kinetic energy). Notice that the combination of (1.47) and (1.49) yields

$$\boldsymbol{p} = c^{-2}E\boldsymbol{u}\,, \tag{1.51}$$

which says that the momentum is a flux of energy (as in electromagnetism, where the density of momentum is proportional to the Poynting vector).

Equation (1.40) can be used to replace $\gamma(u)$ in the energy (1.49); it yields

$$\frac{E}{c} = mc\frac{\mathrm{d}t}{\mathrm{d}\tau}\,. \tag{1.52}$$

Then E is proportional to the ratio of the time $\mathrm{d}t$ measured by frame clocks to the respective proper time of the particle. As stated above, the invariance of $\mathrm{d}\tau$ in (1.47) and (1.52) implies that $(E/c, \boldsymbol{p})$ transforms like $(c\,\mathrm{d}t, \mathrm{d}\boldsymbol{r})$ under Lorentz transformations, i. e.,

$$\begin{aligned} E' &= \gamma(V)(E - c\beta p_x) \\ &= \gamma(V)(E - \boldsymbol{V}\cdot\boldsymbol{p}), \end{aligned} \tag{1.53a}$$

$$p'_x = \gamma(V)(p_x - \beta c^{-1}E)\,, \tag{1.53b}$$

$$p'_y = p_y, \tag{1.53c}$$

$$p'_z = p_z. \tag{1.53d}$$

E^2 and $c^2|\boldsymbol{p}|^2$ combine to yield the square particle mass, an invariant result called the *energy–momentum invariant*

$$\begin{aligned} E^2 - c^2|\boldsymbol{p}|^2 &= m^2c^4\gamma(u)^2 - m^2c^2u^2\gamma(u)^2 \\ &= m^2c^4\left(1-\frac{u^2}{c^2}\right)\gamma(u)^2 = m^2c^4\,. \end{aligned} \tag{1.54}$$

Let us differentiate (1.54) to obtain

$$E\mathrm{d}E = c^2\boldsymbol{p}\cdot\mathrm{d}\boldsymbol{p}\,, \tag{1.55}$$

or, replacing $\boldsymbol{p}$ with (1.51)

$$\mathrm{d}E = \boldsymbol{u}\cdot\mathrm{d}\boldsymbol{p} = \mathrm{d}\boldsymbol{r}\cdot\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t}\,, \tag{1.56}$$

which suggests that the force is associated with $\mathrm{d}\boldsymbol{p}/\mathrm{d}t$. If so, (1.56) would express the equality between the work of the force and the variation of the energy. Notice that $\boldsymbol{F} = \mathrm{d}\boldsymbol{p}/\mathrm{d}t$ implies that the force is not parallel to the acceleration in general, due to the term containing the derivative of $\gamma(u)$. Remarkably, if the work goes to

infinity, then the energy diverges and the velocity u in (1.49) goes to c. In this way, the speed of light is an unreachable limit for the particle.

In electromagnetism, the interaction of a charge with a given external field is described by adding the action (1.46) with the term $S_{\text{int}} = -q \int (\varphi - \boldsymbol{u} \cdot \boldsymbol{A})\, \mathrm{d}t$, where φ and $\boldsymbol{A}$ are the scalar and vector potentials evaluated at the position of the charge. It can be proven that the interaction action S_{int} is Lorentz-invariant, as required in special relativity. The variation of the action $S_{\text{free}} + S_{\text{int}}$ leads to the equation of motion

$$q(\boldsymbol{E} + \boldsymbol{u} \times \boldsymbol{B}) = \frac{\mathrm{d}}{\mathrm{d}t}\,(m\gamma(u)\boldsymbol{u})\,, \tag{1.57}$$

where $\boldsymbol{E} = -\nabla\varphi - \partial\boldsymbol{A}/\partial t$ and $\boldsymbol{B} = \nabla \times \boldsymbol{A}$. In (1.57) we recognize the Lorentz force on the left-hand side and the derivative of the relativistic momentum (1.47) on the right-hand side. In 1908 *Bucherer* [1.41] observed the movement of an electron in an electrostatic field and obtained an incontestable evidence of the validity of the relativistic dynamics expressed in (1.57). If the charge is initially at rest in a uniform static field $\boldsymbol{E}$, then we integrate (1.57) to obtain $(q/m)\, \boldsymbol{E}t = \gamma(u)\boldsymbol{u}$. So, u goes to c when t goes to infinity.

1.5.2 Photons

In 1905 *Einstein* [1.42] stated that the photoelectric effect could be better understood by proposing that light interacts with individual electrons by exchanging packets of energy $h\nu$ (h is Planck's constant and ν is the frequency of light). In this way, the understanding of light–matter interactions required a new concept where light shared characteristics of both wave and corpuscle. In 1917 *Einstein* [1.43] convinced himself that the *quantum of light* should be also endowed with directed momentum, like any particle. The reality of the *photon* was confirmed by *Compton*'s experiment in 1923 [1.44], where the energy–momentum exchange between a photon and a free electron was measured. The energy and momentum of photons traveling along the $\hat{\boldsymbol{n}}$ direction,

$$E_{\text{photon}} = h\nu\,, \quad \boldsymbol{p}_{\text{photon}} = \frac{h\nu}{c}\hat{\boldsymbol{n}}\,, \tag{1.58}$$

are those of a particle having zero mass (1.54) and the speed of light (1.51). Lorentz transformations (1.53) for the energy and the momentum (1.58) become the transformations (1.43) and (1.44) for the frequency and the propagation direction of a ray of light [1.45].

1.5.3 Mass–Energy Equivalence

In relativity, the conservations of momentum and energy cannot be dissociated. While the conservation of momentum comes from the symmetry of the Lagrangian under spatial translations, the conservation of energy results from the symmetry under time translation. However space and time are frame-dependent projections of spacetime. Space and time intermingle under Lorentz transformations. Consequently, the conservation of momentum in all inertial frames requires the conservation of energy and vice versa. This conclusion is evident in the transformations (1.53), where energy and momentum mix under a change of frame; so, the momentum would not be conserved in frame S' if the energy were not conserved in S. In sum, the conserved quantity associated to the symmetry of the Lagrangian under spacetime translations is the total energy–momentum.

In classical mechanics, instead, the transformation of the momentum of the particle does not involve its energy. In fact, if (1.8) is multiplied by the mass, then the transformation $\boldsymbol{p}' = \boldsymbol{p} - m\boldsymbol{V}$ is obtained. Thus, an isolated system of interacting particles conserves the total momenta in all the inertial frames irrespective of what happens with the classical energy. Noticeably, $\Sigma\boldsymbol{p}'$ is conserved whenever $\Sigma\boldsymbol{p}$ is conserved because the total mass Σm is assumed to be a conserved quantity (the classical principle of conservation of mass). This is no longer true in special relativity. For instance, let us consider the plastic collision between two isolated particles of equal mass m. In the *center-of-momentum* frame the (conserved) total momentum vanishes; so the particles have equal and opposite velocities u before the collision. In the collision, the masses stick together and remain at rest. If no energy is released, then the conservation of energy implies

$$2m\gamma(u)c^2 = Mc^2\,, \tag{1.59}$$

where M is the mass of the resulting body. Since $\gamma(u) > 1$, then $M > 2m$; in fact, the resulting body contains the masses of the colliding particles and their kinetic energies. In Einstein's words, *the mass of a body is a measure of its energy-content* [1.46].

In general, the mass (energy at rest) of a composed system includes not only the masses of its constituents but any other internal energy as measured in the center-of-momentum frame. For instance, a deuteron D is constituted by a proton and a neutron. The deuteron mass is lower than the addition of the masses of a free proton

and a free neutron; this evidences a negative binding energy between the constituents. The *mass defect* is $(m_D - m_p - m_n)c^2 = -2.22\,\text{MeV}$. In general, when light nuclides merge into a heavier nuclide (*nuclear fusion*) some energy must be released to conserve the total energy. On the contrary, the mass of a heavy nucleus is larger than the sum of the masses of its constituents. Therefore, also there is a released energy in the *nuclear fission* of heavy nuclei. This dissimilar behavior comes from the fact that the (negative) binding energy per nucleon increases with the mass number for light nuclei but decreases for heavy nuclei (the inversion of the slope happens at a mass number around 60).

The kinetic energy can be used to create particles. For instance, a neutral pion π^0 can be created in a high energy collision between protons p; the reaction is $p + p \to p + p + \pi^0$. This reaction can only occur if a *threshold energy* is reached to give account of the particle created. The neutral pion has energy at rest (mass) of 134.98 MeV; then, in the center-of-momentum frame the pion is created if each colliding proton reaches the kinetic energy of 67.49 MeV. In such a case, all the kinetic energy is used to create the pion; the products remain at rest, since no kinetic energy is left for the products, and the total momentum is conserved. Therefore, the threshold energy of the reaction in the center-of-momentum frame is equal to the energy at rest of the products: $E_{\text{threshold}} = 2m_p c^2 + m_\pi^0 c^2 = 1876.54\,\text{MeV} + 134.98\,\text{MeV}$. In this case, the energy balance is (the particles are approximately free before and after the reaction)

$$\begin{aligned} 2m_p\gamma(u_p)c^2 &= 2m_p c^2 + m_{\pi^0}c^2 \\ \Rightarrow \quad \gamma(u_p) &= 1 + \frac{m_{\pi^0}}{2m_p} = 1 + \frac{134.98}{1876.54} \\ &= 1.072\,, \end{aligned} \tag{1.60}$$

which means that the velocity of the colliding protons in the center-of-momentum frame is $u_p = 0.36c$. In another frame, the threshold energy is higher because the products must keep some kinetic energy to conserve the (non-null) total momentum. We can use (1.53) to transform the total energy-momentum of the system (since the transformations are linear, they can be used to transform a sum of energies and momenta). In the center-of-momentum frame the total momentum is zero; then (1.53a) says that $E'_{\text{threshold}} = \gamma(V)E_{\text{threshold}}$. For instance, in the *laboratory frame* where one of the colliding protons is at rest (i. e., $\gamma(V) = \gamma(u_p)$) it is $E'_{\text{threshold}} = 1.072E_{\text{threshold}}$; deducting the masses of projectile and target, we obtain that the reaction is feasible if the projectile reaches the kinetic energy of $T'_{\text{threshold}} = E'_{\text{threshold}} - 2m_p c^2 = 279.67\,\text{MeV}$.

The previous example is a case of inelastic collision. A collision is called *elastic* if the particles keep their identities. Thus, the masses (energies at rest) before and after the collision are the same; so, the conservation of the energy of the colliding free particles is equivalent to the conservation of the total kinetic energy.

The interaction among charged particles can result in the release of electromagnetic radiation. In such cases the radiation enters the energy-momentum balance in the form of photons. For instance, a pair electron-positron annihilates to give two photons (the positron is the anti-particle of the electron; they have equal mass but opposite charge). In the center-of-momentum frame, the photons have equal frequency and opposite directions to conserve the total momentum (notice that two photons at least are needed to conserve the momentum). If u_e is the velocity of both particles in the center-of-momentum frame, then the energy balance is

$$2m_e\gamma(u_e)c^2 = 2h\nu\,. \tag{1.61}$$

Conversely, two photons can create a pair electron–positron. In this case, the threshold energy is equal to the mass of two electrons. So the minimum frequency to create the pair in the center-of-momentum frame is given by

$$\begin{aligned} 2h\nu_{\text{min}} &= 2m_e c^2 \\ \Rightarrow \quad \nu_{\text{min}} &= \frac{m_e c^2}{h} = \frac{0.511\,\text{MeV}}{4.14\times10^{-21}\,\text{MeV s}} \\ &= 1.23\times10^{20}\,\text{s}^{-1}\,, \end{aligned} \tag{1.62}$$

which is a frequency in the gamma-ray range of the electromagnetic spectrum.

Compton Effect

In 1923 Compton measured the scattering of x-rays by electrons in graphite. x-ray photons have energies much larger than the electron bound energies. So, the phenomenon can be studied as the elastic collision between a photon and a free electron. In the frame where the electron is initially at rest, its final momentum and energy are

$$\begin{aligned} E_e &= h\nu_i - h\nu_f + m_e c^2\,, \\ \boldsymbol{p}_e &= h\nu_i c^{-1}\hat{\boldsymbol{n}}_i - h\nu_f c^{-1}\hat{\boldsymbol{n}}_f\,, \end{aligned} \tag{1.63}$$

as results from compensating the changes of momentum and energy suffered by photon and electron (in (1.63) the labels i and f allude to the initial and final states of the photon). The replacement of these values in the electron energy–momentum invariant (1.54) yields

$$\begin{aligned} m_e^2c^4 &= E_e^2 - p_e^2c^2 \\ &= m_e^2c^4 + 2h\left(\nu_i m_e c^2 - h\nu_i\nu_f - m_e c^2\nu_f\right. \\ &\qquad \left. + h\nu_i\nu_f \hat{\boldsymbol{n}}_i \cdot \hat{\boldsymbol{n}}_f\right) . \end{aligned} \tag{1.64}$$

Equation (1.64) contains the relation between the ingoing and outgoing photons. Let us call φ the angle between the initial and final directions of propagation: $\hat{\boldsymbol{n}}_i \cdot \hat{\boldsymbol{n}}_f = \cos\varphi$. Then

$$\begin{aligned} \frac{1}{h\nu_f} - \frac{1}{h\nu_i} = \frac{1}{m_e c^2}(1-\cos\varphi) \quad \text{or} \quad \lambda_f - \lambda_i \\ = \frac{h}{m_e c}(1-\cos\varphi) . \end{aligned} \tag{1.65}$$

The quantity $\lambda_C \equiv h/(m_e c) = 0.00243$ nm is the *Compton wavelength* of the electron. Equation (1.65) says that the photon suffers a significant change only if its wavelength is comparable to or smaller than the electron Compton wavelength (i. e., its energy is comparable to or larger than $m_e c^2$).

1.5.4 Interactions *at a Distance*

Interactions at a distance are allowed in classical mechanics; they are described by potential energies depending on the distances between particles, which automatically give equal and opposite interaction forces accomplishing Newton's third law. Thus, although the interaction forces change the momenta of the particles, these changes cancel out by pairs at each instant; so the total momentum of an isolated system of interacting particles is conserved. Noticeably, the statement of Newton's third law cannot be translated to special relativity, because the *simultaneous* cancelation at a distance does not have an absolute meaning. In particular, an interaction potential energy depending on the (non-Lorentz-invariant) distance between particles makes no sense in relativity. Remarkably, in electromagnetism the charges do not interact through such a potential (apart from the static case). Instead, the interaction at a distance is substituted for the *local* interaction between a charge and the surrounding electromagnetic field. This local interaction entails the exchange of energy and momentum between charge and field. The electromagnetic field carries momentum and energy, which can be (partially) transferred to another charge at another place. So, the isolated system conserving the total momentum and energy is composed by the charges *and* the electromagnetic field. Conservation laws are local in relativity. The action governing an isolated system of charges and electromagnetic field is the sum of the actions S_{free} of the charges, the actions S_{int} describing the local interaction of each charge with the field at the place of the charge (Sect. 1.5.1), and the invariant action of the electromagnetic field $S_{field} = \varepsilon_o/2\int(\boldsymbol{E}^2 - c^2\boldsymbol{B}^2)\,d^3\boldsymbol{x}\,dt$.

1.6 Conclusion

As a theory about the structure of the spacetime, special relativity is a framework to build theories in physics: the laws governing any physical phenomenon must be derived from Lorentz invariant functional actions. In this way, the dynamical equations will accomplish the principle of relativity under Lorentz transformations.

This requirement is enlightened in the *covariant* formulation to be developed in the next chapters. Certainly, Maxwell's electromagnetism is a theory that has the proper behavior under Lorentz transformations. Also the field theories describing subatomic interactions are built under relativistic criteria. What about the theory of gravity? In classical physics, gravity is a universal force proportional to the mass. The identity between the *gravitational mass* – the mass that measures the strength of the gravitational interaction – and the *inertial mass* – the mass in (1.11) – causes the motion of a *freely-gravitating* particle to be independent of its mass; it just depends on the initial conditions. Einstein realized that this fact opened up the possibility of considering gravity not as a force but as the geometry of spacetime: the motion of a freely gravitating particle would be the consequence of the geometry of the spacetime. Special relativity had revised the belief in the invariance of lengths and times, but it still assumed that the space was endowed with a frozen Euclid's flat geometry (which leads to the Pythagoras' theorem we used in (1.27)). Einstein took a big

step ahead to think that geometry could be a dynamical variable determined by the distribution of matter and energy. Thus, Newton's thought that matter is the origin of gravitational forces was replaced by Einstein's idea that the energy–momentum distribution determines the way of measuring spacetime. In general relativity, geometry is governed by dynamical equations – the Einstein equations – fed by the energy and momentum located in the spacetime; special relativity's geometry is just the geometry of an empty spacetime. In general relativity, the freely gravitating test particles describe *geodesics* of the spacetime geometry; this is what a planet does when orbiting a star. Moreover, when a photon ascends a gravitational field, its frequency diminishes because clocks go faster when the gravitational potential increases ((1.40) is no longer valid). The GPS system takes into account this effect of gravity on the running of clocks to reach its highest performance. So, the photon loses energy while ascending a gravitational potential. This implies that its capacity of creating mass decreases; but the so created mass is compensated for by a larger *potential energy*. In general relativity the spacetime geometry can evolve; thus we can interpret the cosmological data in the context of an expanding universe. In sum, 10 years after the birth of special relativity, the concepts of space and time underwent a new fundamental revision to tackle the relativistic formulation of gravitational phenomena: Einstein's general relativity was born.

References

1.1 I. Newton: *Philosophiæ Naturalis Principia Mathematica* (Joseph Streater, London 1687)
1.2 G. Galilei: *Discorsi e Dimostrazioni Matematiche, intorno à due nuove scienze* (Elsevirii, Leiden 1638), Third Day, Section 243
1.3 P. Gassendi: *De motu impresso a motore translato* (Louis de Heuqueville, Paris 1642)
1.4 R. Descartes: *Principia Philosophiae* (Louis Elzevir, Amsterdam 1644), Part II, Section 37
1.5 H.G. Alexander (Ed.): *The Leibniz–Clarke Correspondence Together with Extracts from Newton's Principia and Opticks* (Manchester Univ. Press, Manchester 1998)
1.6 O.C. Rømer: Démonstration touchant le mouvement de la lumière trouvé par M. Römer de l'Académie Royale des Sciences, J. Sçavans, 233-236 (1676)
1.7 L. Bobis, J. Lequeux: Cassini, Rømer and the velocity of light, J. Astron. Hist. Herit. **11**(2), 97–105 (2008)
1.8 J. Bradley: A letter from the Reverend Mr. James Bradley ... to Dr. Edmond Halley ... giving an account of a new discovered motion of the fixed stars, Philos. Trans. R. Soc. **35**, 637–661 (1728)
1.9 A.B. Stewart: The discovery of stellar aberration, Sci. Am. **210**(3), 100–108 (1964)
1.10 H.L. Fizeau: Sur une expérience relative à la vitesse de propagation de la lumière, C. R. Acad. Sci. Paris **29**, 90–92 (1849)
1.11 A.J. Fresnel: Mémoire sur la diffraction de la lumière, Mém. Acad. Sci. **5**, 339–475 (1821), 1822
1.12 C. Huygens: *Traité de la Lumière* (Pierre van der Aa, Leiden 1690)
1.13 L. Foucault: Méthode générale pour mesurer la vitesse de la lumière dans l'air et les milieux transparents. Vitesses relatives de la lumière dans l'air et dans l'eau. Projet d'expérience sur la vitesse de propagation du calorique rayonnant, C. R. Acad. Sci. Paris **30**(18), 551–560 (1850)
1.14 J.C. Maxwell: *A Treatise on Electricity and Magnetism* (Clarendon, Oxford 1873)
1.15 M. Hoek: Détermination de la vitesse avec laquelle est entraîné une onde lumineuse traversant un milieu en movement, Arch. Néerl. Sci. **3**, 180–185 (1868)
1.16 G.B. Airy: On a supposed alteration in the amount of astronomical aberration of light, produced by the passage of the light trough a considerable thickness of refracting medium, Proc. R. Soc. **20**, 35–39 (1871)
1.17 D.F.J. Arago: Mémoire sur la vitesse de la lumière, lu à la première Classe de l'Institut, le 10 décembre 1810, C. R. Acad. Sci. Paris **36**(2), 38–49 (1853)
1.18 A.J. Fresnel: Lettre de M Fresnel à M Arago, sur l'influence du mouvement terrestre dans quelques phénomènes d'optique, Ann. Chim. Phys. **9**, 57–66 (1818)
1.19 H.L. Fizeau: Sur les hypotheses relatives à l'éther lumineux, et sur une expérience qui paraît démontrer que le mouvement des corps change la vitesse avec laquelle la lumière se propage dans leur intérieur, C. R. Acad. Sci. Paris **33**(15), 349–355 (1851)
1.20 R. Ferraro: *Einstein's Space-Time: An Introduction to Special and General Relativity* (Springer, New York 2007)
1.21 R. Ferraro, D.M. Sforza: Arago (1810): the first experimental result against the ether, Eur. J. Phys. **26**, 195–204 (2005)
1.22 E. Hecht: *Optics* (Addison-Wesley, Reading 2002)
1.23 A.A. Michelson, E.W. Morley: On the relative motion of the Earth and the luminiferous ether, Am. J. Sci. **34**, 333–345 (1887)
1.24 A.A. Michelson, E.W. Morley: On the relative motion of the Earth and the luminiferous ether, Philos. Mag. **24**, 449–463 (1887)

1.25 L.S. Swenson: The Michelson–Morley–Miller experiments before and after 1905, J. Hist. Astron. **1**, 56–78 (1970)
1.26 O.J. Lodge: Aberration problems. A Discussion concerning the motion of the ether near the earth, and concerning the connexion between ether and gross matter; with some new experiments, Philos. Trans. R. Soc. A **184**, 727–804 (1893)
1.27 W. Ritz: Recherches Critiques sur l'Électrodynamique Générale, Ann. Chim. Phys. **13**, 145–275 (1908)
1.28 G.C. Babcock, T.G. Bergman: Determination of the constancy of the speed of light, J. Opt. Soc. Am. **54**, 147–150 (1964)
1.29 J.G. Fox: Evidence against emission theories, Am. J. Phys. **33**, 1–17 (1965)
1.30 A.A. Martínez: Ritz, Einstein, and the emission hypothesis, Phys. Perspect. **6**, 4–28 (2004)
1.31 H.A. Lorentz: De relatieve beweging van de Aarde en den Aether, Verh. K. Akad. Wet. **1**, 74–79 (1892)
1.32 G.F. FitzGerald: The ether and the Earth's atmosphere, Science **13**, 390 (1889)
1.33 A. Einstein: Zur Elektrodynamik bewegter Körper, Ann. Phys. **17**, 891–921 (1905)
1.34 B. Rossi, D.B. Hall: Variation of the rate of decay of mesotrons with momentum, Phys. Rev. **59**, 223–228 (1941)
1.35 H.A. Lorentz: Versuch einer Theorie der electrischen und optischen Erscheinungen in bewegten Körpern, Verh. K. Akad. Wet. **7**, 507–522 (1899), translation: Théorie simplifieé des phénomènes electriques et optiques dans les corps en movement, Proc. Sect. Sci. K. Akad. Wet. Amst. **1**, 427-442
1.36 H.A. Lorentz: K. Akad. Wet. **12**, 986 (1904)
1.37 J. Larmor: *Æther and Matter* (Cambridge Univ. Press, Cambridge 1900)
1.38 W. Voigt: Ueber das Doppler'sche Princip, Goett. Ges. Wiss. Nachr. **2**, 41–51 (1887)
1.39 J.H. Poincaré: Sur la dynamique de l'électron, C. R. Acad. Sci. Paris **140**, 1504–1508 (1905)
1.40 H.E. Ives, G.R. Stilwell: An experimental study of the rate of a moving atomic clock, J. Opt. Soc. Am. **28**, 215–219 (1938)
1.41 A.H. Bucherer: Messungen an Becquerelstrahlen. Die experimentelle Bestätigung der Lorentz-Einsteinschen Theorie, Phys. Z. **9**, 755–762 (1908)
1.42 A. Einstein: Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt, Ann. Phys. **17**, 132–148 (1905)
1.43 A. Einstein: Zur Quantentheorie der Strahlung, Phys. Z. **18**, 121–128 (1917)
1.44 A.H. Compton: A quantum theory of the scattering of x-rays by light elements, Phys. Rev. **21**, 483–502 (1923)
1.45 A. Cassini, M.L. Levinas: La hipótesis del cuanto de luz y la relatividad especial ¿Por qué Einstein no las relacionó en 1905?, Sci. Stud. **5**(4), 425–452 (2007)
1.46 A. Einstein: Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig?, Ann. Phys. **18**, 639–641 (1905)

2. The Historical Origins of Spacetime

Scott Walter

The idea of spacetime investigated in this chapter, with a view toward understanding its immediate sources and development, is the one formulated and proposed by Hermann Minkowski in 1908. Until recently, the principle source used to form historical narratives of *Minkowski*'s discovery of spacetime has been *Minkowski*'s own discovery account, outlined in the lecture he delivered in Cologne, entitled *Space and time* [2.1]. Minkowski's lecture is usually considered as a *bona fide* first-person narrative of lived events. According to this received view, spacetime was a natural outgrowth of Felix Klein's successful project to promote the study of geometries via their characteristic groups of transformations. Or as Minkowski expressed the same basic thought himself, the theory of relativity discovered by physicists in 1905 could just as well have been proposed by some late-nineteenth-century mathematician, by simply reflecting upon the groups of transformations that left invariant the form of the equation of a propagating light wave. Minkowski's publications and research notes provide a contrasting picture of the discovery of spacetime, in which group theory plays no direct part. In order to relate the steps of Minkowski's discovery, we begin with an account of Poincaré's theory of gravitation, where Minkowski found some of the germs of spacetime. Poincaré's geometric interpretation of the Lorentz transformation is examined, along with his reasons for not pursuing a four-dimensional vector calculus. In the second section, Minkowski's discovery and presentation of the notion of a world line in spacetime is presented. In the third and final section, Poincaré's and Minkowski's diagrammatic interpretations of the Lorentz transformation are compared.

2.1 Poincaré's Theory of Gravitation

In the month of May, 1905, Henri Poincaré (1854–1912) wrote to his Dutch colleague H. A. Lorentz (1853–1928) to apologize for missing the latter's lecture in Paris, and also to communicate his latest discovery, which was related to *Lorentz*'s recent paper [2.2] on electromagnetic phenomena in frames moving with sublight velocity [2.3, §38.3]. In [2.2], *Lorentz* had shown that the form of the fundamental equations of his theory of electrons is invariant with respect to the coordinate transformations

$$x' = \gamma\ell x\,, \quad y' = \ell y\,, \quad z' = \ell z\,,$$
$$t' = \frac{\ell}{\gamma}t - \beta\ell\frac{v}{c^2}x\,, \tag{2.1}$$

where

$$\gamma = 1/\sqrt{1-v^2/c^2}\,,$$
$$\ell = f(v),\ \ell = 1 \text{ for } v = 0\,,$$
$$c = \text{vacuum speed of light}\,.$$

The latter transformation was understood to compose with a transformation later known as a *Galilei* transformation: $x'' = x' - vt'$, $t'' = t'$. (Both here and elsewhere in this chapter, original notation is modified for ease of reading.)

The essence of Poincaré's discovery in May 1905, communicated in subsequent letters to Lorentz, was that the coordinate transformations employed by Lorentz

form a group, provided that the factor ℓ is set to unity. Poincaré performed the composition of the two transformations to obtain a single transformation, which he called the *Lorentz* transformation

$$x' = \gamma(x - vt)\,, \quad y' = y\,, \quad z' = z\,, \\ t' = \gamma\left(t - v\frac{x}{c^2}\right)\,. \tag{2.2}$$

In his letters to Lorentz, Poincaré noted that while he had concocted an electron model that was both stable and relativistic, in the new theory he was unable to preserve the *unity of time*, i. e., a definition of duration valid in both the ether and in moving frames.

The details of *Poincaré*'s theory [2.4] were published in January, 1906, by which time *Einstein* had published his own theory of relativity [2.5], which employed the unified form of the Lorentz transformation (2.2) and vigorously embraced the relativity of space and time with respect to inertial frames of motion. The final section of Poincaré's memoir is devoted to a topic he had neglected to broach with Lorentz, and that Einstein had neglected altogether: gravitation.

If the principle of relativity was to be universally valid, Poincaré reasoned, then Newton's law of gravitation would have to be modified. An adept of the group-theoretical understanding of geometry since his discovery of what he called *Fuchsian* functions in 1880 [2.6], Poincaré realized that a Lorentz transformation may be construed as a rotation about the origin of coordinates in a four-dimensional vector space with three real axes and one imaginary axis, preserving the sum of squares

$$x'^2 + y'^2 + z'^2 - t'^2 = x^2 + y^2 + z^2 - t^2\,, \tag{2.3}$$

where Poincaré set $c = 1$. Employing the substitution $u = t\sqrt{-1}$, and drawing on a method promoted by *Lie* and *Scheffers* in the early 1890s [2.7], Poincaré identified a series of quantities that are invariant with respect to the Lorentz group. These quantities were meant to be the fundamental building blocks of a Lorentz-covariant family of laws of gravitational attraction. Neglecting a possible dependence on acceleration, and assuming that the propagation velocity of gravitation is the same as that of light in empty space, Poincaré identified a pair of laws, one vaguely Newtonian, the other vaguely Maxwellian, which he expressed in the form of what would later be called four-vectors.

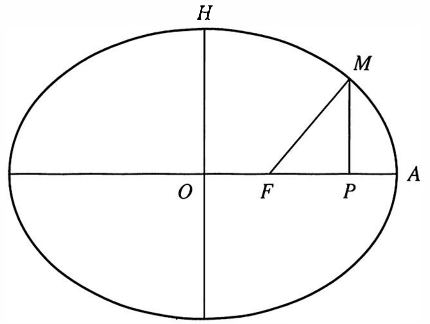
In the course of his work on Lorentz-covariant gravitation, Poincaré defined several quadruples formally equivalent to four-vectors, including definitions of radius, velocity, force, and force density. The signs of Poincaré's invariants suggest that when he formed them, he did *not* consider them to be scalar products of four-vectors. This state of affairs led at least one contemporary observer to conclude – in the wake of Minkowski's contributions – that Poincaré had simply miscalculated one of his Lorentz invariants [2.8, pp. 203; 238].

Poincaré's four-dimensional vector space attracted little attention at first, except from the vectorist Roberto Marcolongo (1862–1945), Professor of Mathematical Physics in Messina. Redefining Poincaré's temporal coordinate as $u = -t\sqrt{-1}$, *Marcolongo* introduced four-vector definitions of current and potential, which enabled him to express the Lorentz-covariance of the equations of electrodynamics in matricial form [2.9]. Largely ignored at the time, Marcolongo's paper nonetheless broke new ground in applying Poincaré's four-dimensional approach to the laws of electrodynamics.

Marcolongo was one of many ardent vectorists active in the first decade of the twentieth century, when vector methods effectively sidelined the rival quaternionic approaches [2.10, p. 259]. More and more theorists recognized the advantages of vector analysis and also of a unified vector notation for mathematical physics. The pages of the leading journal of theoretical physics, the *Annalen der Physik*, edited by Paul Drude until his suicide in 1906, then by Max Planck and Willy Wien, bear witness to this evolution. Even in the pages of the *Annalen der Physik*, however, notation was far from standardized, leading several theorists to deplore the field's babel of symbolic expressions.

Among the theorists who regretted the multiplication of systems of notation was Poincaré, who employed ordinary vectors in his own teaching and publications on electrodynamics, while ignoring the notational innovations of Lorentz and others. In particular, Poincaré saw no future for a four-dimensional vector calculus. Expressing physical laws by means of such a calculus, he wrote in 1907, would entail *much trouble for little profit* [2.11, p. 438].

This was not a dogmatic view, and in fact, some years later he acknowledged the value of a four-dimensional approach in theoretical physics [2.12, p. 210]. He was already convinced that there was a place for $(3+n)$-dimensional geometries at the university. As *Poincaré* observed in the paper Gaston Darboux read in his stead at the International Congress of Mathemati-



Lorentz contraction	$\gamma = 1/\sqrt{1 - v^2/c^2}$
Semi-major axis	$a = OA = \gamma ct$
Semi-minor axis	$b = OH = ct$
Eccentricity	$e = \sqrt{1 - b^2/a^2} = v/c$
Focal length	$f = OF = \gamma vt$
Light path	$p = FM$
Apparent displacement	$x' = FP$

Fig. 2.1 Poincaré's light ellipse, after Henri Vergne, 1906–1907. Labels H and A are added for clarity

cians in Rome, in April, 1908, university students were no longer taken aback by geometries with *more than three dimensions* [2.13, p. 938].

Relativity theory, however, was another matter for Poincaré. Recently-rediscovered manuscript notes by Henri Vergne of Poincaré's lectures on relativity theory in 1906–1907 reveal that Poincaré introduced his students to the Lorentz group and taught them how to form Lorentz-invariant quantities with real coordinates. He also taught his students that the sum of squares (2.3) is invariant with respect to the transformations of the Lorentz group. Curiously, Poincaré did not teach his students that a Lorentz transformation corresponded to a rotation about the origin in a four-dimensional vector space with one imaginary coordinate. He also neglected to show his students the handful of four-vectors he had defined in the summer of 1905. Apparently for Poincaré, knowledge of the Lorentz group and the formation of Lorentz-invariant quantities was all that was needed for the physics of relativity. In other words, Poincaré acted as if one could do without an interpretation of the Lorentz transformation in four-dimensional geometry.

If four-dimensional geometry was superfluous to interpretation of the Lorentz transformation, the same was not true for plane geometry. Evidence of this view is found in Vergne's notes, which feature a curious figure that we will call a light ellipse, redrawn here as Fig. 2.1. Poincaré's light ellipse is given to be the meridional section of an ellipsoid of rotation representing the locus of a spherical light pulse at an instant of time. It works as follows: an observer at rest with respect to the ether measures the radius of a spherical light pulse at an instant of absolute time t (as determined by clocks at rest with respect to the ether). The observer measures the light pulse radius with measuring rods in uniform motion of velocity v. These flying rods are Lorentz-contracted, while the light wave is assumed to propagate spherically in the ether. Consequently, for Poincaré, the form of a spherical light pulse measured in this fashion is that of an ellipsoid of rotation, elongated in the direction of motion of the flying rods. (A derivation of the equation of Poincaré's light ellipse is provided along these lines in [2.14].)

The light ellipse originally concerned ether-fixed observers measuring a locus of light with clocks at absolute rest, and rods in motion. Notably, in his first discussion of the light ellipse, Poincaré neglected to consider the point of view of observers in motion with respect to the ether. In particular, Poincaré's graphical model of light propagation does not display relativity of simultaneity for inertial observers, since it represents a single frame of motion. Nonetheless, Poincaré's light ellipse was applicable to the case of observers in uniform motion, as he showed himself in 1909. In this case, the radius vector of the light ellipse represents the light-pulse radius at an instant of *apparent* time t', as determined by comoving, light-synchronized clocks, and comoving rods corrected for Lorentz-contraction. Such an interpretation implies that clock rates depend on frame velocity, as *Einstein* recognized in 1905 in consequence of his kinematic assumptions about ideal rods and clocks [2.5, p. 904], and which *Poincaré* acknowledged in a lecture in Göttingen on 28 April, 1909, as an effect epistemically akin to Lorentz-contraction, induced by clock motion with respect to the ether [2.15, p. 55].

Beginning in August 1909, *Poincaré* repurposed his light ellipse diagram to account for the dilation of periods of ideal clocks in motion with respect to the ether [2.16, p. 174]. This sequence of events raises the question of what led Poincaré to embrace the notion of time deformation in moving frames, and to repurpose his light ellipse? He did not say, but there is a plausible explanation at hand, which we will return to later, as it rests on events in the history of relativity from 1907 to 1908 to be discussed in the next section.

2.2 Minkowski's Path to Spacetime

From the summer of 1905 to the fall of 1908, the theory of relativity was reputed to be inconsistent with the observed deflection of β-rays by electric and magnetic fields. In view of experimental results published by Walter Kaufmann (1871–1947), Lorentz wrote in despair to Poincaré on 8 March, 1906 in hopes that the Frenchman would find a way to save his theory. As far as Lorentz was concerned, he was *at the end of [his] Latin* [2.17, p. 334].

Apparently, Poincaré saw no way around Kaufmann's results, either. However, by the end of 1908, the outlook for relativity theory had changed for the better, due in part to new experiments performed by A. H. Bucherer (1863–1927), which tended to confirm the predictions of relativity theory. The outlook for the latter theory was also enhanced by the contributions of a mathematician in Göttingen, Hermann Minkowski (1864–1909).

Minkowski's path to theoretical physics was a meandering one, that began in earnest during his student days in Berlin, where he heard lectures by Hermann Helmholtz, Gustav Kirchhoff, Carl Runge, and Woldemar Voigt. There followed a dissertation in Königsberg on quadratic forms, and *Habilitation* in Bonn on a related topic in 1887 [2.18]. While in Bonn, Minkowski frequented Heinrich Hertz's laboratory beginning in December, 1890, when it was teeming with young physicists eager to master techniques for the study of electromagnetic wave phenomena. Minkowski left Bonn for a position at the University of Königsberg, where he taught mathematics until 1896, and then moved to Switzerland, where he joined his former teacher Adolf Hurwitz (1859–1919) on the faculty of Zürich Polytechnic. In Zürich he taught courses in mathematics and mathematical physics to undergraduates including Walther Ritz (1878–1909), Marcel Grossmann (1878–1936), and Albert Einstein (1879–1955). In 1902, Minkowski accepted the offer to take up a new chair in mathematics created for him in Göttingen at the request of his good friend, David Hilbert (1862–1943) [2.19, p. 436].

Minkowski's arrival in Göttingen comforted the university's premier position in mathematical research in Germany. His mathematical credentials were well-established following the publication, in 1896, of the seminal *Geometry of numbers* [2.20]. During his first 2 years in Göttingen, Minkowski continued to publish in number theory and to teach pure mathematics. With Hilbert, who had taken an interest in questions of mathematical physics in the 1890s, Minkowski codirected a pair of seminars on stability and mechanics [2.21].

It was quite unusual at the time for Continental mathematicians to pursue research in theoretical physics. Arguably, Poincaré was the exception that proved the rule, in that no other scientist displayed comparable mastery of research in both mathematics and theoretical physics. In Germany, apart from Carl Neumann, mathematicians left physics to the physicists. With the construction in Germany of 12 new physical institutes between 1870 and 1899, there emerged a professional niche for individuals trained in both mathematical physics and experimental physics, which very few mathematicians chose to enter. This *institutional revolution* in German physics [2.22] gave rise to a new breed of physicist: the *theoretical* physicist [2.23].

In the summer of 1905, Minkowski and Hilbert codirected a third seminar in mathematical physics, convinced that only higher mathematics could solve the problems then facing physicists, and with Poincaré's 14 volumes of Sorbonne lectures on mathematical physics serving as an example. This time they delved into a branch of physics new to both of them: electron theory. Their seminar was an occasion for them to acquaint themselves, their colleagues Emil Wiechert and Gustav Herglotz, and students including Max Laue and Max Born, with recent research in electron theory. From all accounts, the seminar succeeded in familiarizing its participants with the state of the art in electron theory, although the syllabus did not feature the most recent contributions from Lorentz and Poincaré [2.24]. In particular, according to Born's distant recollections of the seminar, Minkowski *occasionally hinted* of his engagement with the Lorentz transformation and he conveyed an *inkling* of the results he would publish in 1908 [2.25].

The immediate consequence of the electron-theory seminar for Minkowski was a new interest in a related, and quite-puzzling topic in theoretical physics: black-body radiation. Minkowski gave two lectures on heat radiation in 1906 and offered a lecture course in this subject during the summer semester of 1907. According to Minkowski's class notes, he referred to *Max Planck*'s contribution to the foundations of relativistic thermodynamics [2.26], which praised Einstein's formulation of a general approach to the principle of relativity for ponderable systems. In fact, Minkowski had little time to assimilate Planck's findings (communicated on 13 June,

1907) and communicate them to his students. This may explain why his lecture notes cover only nonrelativistic approaches to heat radiation.

By the fall of 1907, Minkowski had come to realize some important consequences of relativity theory not only for thermodynamics, but for all of physics. On 9 October, he wrote to Einstein, requesting an offprint of his first paper on relativity, which was the one cited by Planck [2.27, Doc. 62]. Less than a month later, on 5 November, 1907, Minkowski delivered a lecture to the Göttingen Mathematical Society, the subject of which was described succinctly as *On the principle of relativity in electrodynamics: a new form of the equations of electrodynamics* [2.25].

The lecture before the mathematical society was the occasion for *Minkowski* to unveil a new research program: to reformulate the laws of physics in four-dimensional terms, based on the Lorentz-invariance of the quadratic form $x^2+y^2+z^2-c^2t^2$, where x, y, z, are rectangular space coordinates, fixed *in ether*, t is time, and c is the vacuum speed of light [2.28, p. 374]. Progress toward the achievement of such a reformulation had been realized by Poincaré's relativistic reformulation of the law of gravitation in terms of Lorentz-invariant quantities expressed in the form of four-vectors, as mentioned above.

Poincaré's formal contribution was duly acknowledged by Minkowski, who intended to go beyond what the Frenchman had accomplished in 1905. He also intended to go beyond what Poincaré had considered to be desirable, with respect to the application of geometric reasoning in the physical sciences. *Poincaré*, we recall, had famously predicted that Euclidean geometry would forever remain the most convenient one for physics [2.11, p. 45]. Poincaré's prediction stemmed in part from his doctrine of physical space, according to which the question of the geometry of phenomenal space cannot be decided on empirical grounds. In fact, few of Poincaré's contemporaries in the physical and mathematical sciences agreed with his doctrine [2.29].

Euclidean geometry was to be discarded in favor of a certain four-dimensional manifold, and not just any manifold, but a *non-Euclidean* manifold. The reason for this was metaphysical, in that for *Minkowski*, the phenomenal world was not Euclidean, but non-Euclidean and four-dimensional [2.28, p. 372]:

> *The world in space and time is, in a certain sense, a four-dimensional non-Euclidean manifold.*

Explaining this enigmatic proposition would take up the rest of Minkowski's lecture.

To begin with, Minkowski discussed neither space, time, manifolds, or non-Euclidean geometry, but vectors. Borrowing Poincaré's definitions of radius and force density, and adding (like Marcolongo before him) expressions for four-current density, ϱ, and four-potential, ψ, Minkowski expressed Maxwell's vacuum equations in the compact form

$$\Box\psi_j = -\varrho_j \quad (j=1,2,3,4)\,, \tag{2.4}$$

where $\Box$ is the d'Alembertian operator. According to Minkowski, no one had realized before that the equations of electrodynamics could be written so succinctly, *not even Poincaré* (cf. [2.30]). Apparently, Minkowski had not noticed Marcolongo's paper, mentioned above.

The next mathematical object that Minkowski introduced was a real step forward and was soon acknowledged as such by physicists. This is what Minkowski called a *Traktor*: a six-component object later called a *six-vector*, and more recently, an antisymmetric rank-2 tensor. Minkowski defined the Traktor's six components via his four-vector potential, using a two-index notation: $\psi_{jk} = \partial\psi_k/\partial x_j - \partial\psi_j/\partial x_k$, noting the antisymmetry relation $\psi_{kj} = -\psi_{jk}$, and zeros along the diagonal $\psi_{jj} = 0$, such that the components ψ_{14}, ψ_{24}, ψ_{34}, ψ_{23}, ψ_{31}, ψ_{12} match the field quantities $-iE_x$, $-iE_y$, $-iE_z$, B_x, B_y, B_z. To express the source equations, Minkowski introduced a *Polarisationstraktor* p

$$\frac{\partial p_{1j}}{\partial x_1} + \frac{\partial p_{2j}}{\partial x_2} + \frac{\partial p_{3j}}{\partial x_3} + \frac{\partial p_{4j}}{\partial x_4} = \sigma_j - \varrho_j\,, \tag{2.5}$$

where σ is the four-current density for matter.

Up to this point in his lecture, Minkowski had presented a new and valuable mathematical object, the antisymmetric rank-2 tensor. He had yet to reveal the sense in which the world is a *four-dimensional non-Euclidean manifold*. His argument proceeded as follows. The tip of a four-dimensional velocity vector w_1, w_2, w_3, w_4, *Minkowski* explained [2.28, p. 373],

> *is always a point on the surface*
>
> $$w_1^2 + w_2^2 + w_3^2 + w_4^2 = -1\,, \tag{2.6}$$
>
> *or if you prefer, on*
>
> $$t^2 - x^2 - y^2 - z^2 = 1\,, \tag{2.7}$$
>
> *and represents both the four-dimensional vector from the origin to this point, and null velocity, or rest, being a genuine vector of this sort. Non-Euclidean geometry, of which I spoke earlier in an*

imprecise fashion, now unfolds for these velocity vectors.

These two surfaces, a pseudo-hypersphere of unit imaginary radius (2.6), and its real counterpart, the two-sheeted unit hyperboloid (2.7), give rise to well-known models of hyperbolic space, popularized by *Helmholtz* in the late nineteenth century [2.31, Vol. 2]. The upper sheet ($t > 0$) of the unit hyperboloid (2.7) models hyperbolic geometry; for details, see [2.32].

The conjugate diameters of the hyperboloid (2.7) give rise to a geometric image of the Lorentz transformation. Any point on (2.7) can be considered to be at rest, i. e., it may be taken to lie on a t-diameter, as shown in Fig. 2.2. This change of axes corresponds to an orthogonal transformation of the time and space coordinates, which is a Lorentz transformation (letting $c = 1$). In other words, the three-dimensional hyperboloid (2.7) embedded in four-dimensional pseudo-Euclidean space affords an interpretation of the Lorentz transformation.

Although Minkowski did not spell out his geometric interpretation, he probably recognized that a displacement on the hypersurface (2.7) corresponds to a rotation ψ about the origin, such that frame velocity v is described by a hyperbolic function, $v = \tanh \psi$. However, he did not yet realize that his hypersurfaces represent the set of events occurring at coordinate time $t' = 1$ of all inertial observers, the world lines of whom pass through the origin of coordinates (with a common origin of time). According to (2.7), this time is imaginary, a fact which may have obscured the latter interpretation.

How do we know that Minkowski was still unaware of world lines in spacetime? Inspection of Minkowski's definition of four-velocity vectors reveals an error, which is both trivial and interesting: trivial from a mathematical standpoint, and interesting for what it says about his knowledge of the structure of spacetime, and the progress he had realized toward his goal of replacing the Euclidean geometry of phenomenal space with the geometry of a four-dimensional non-Euclidean manifold.

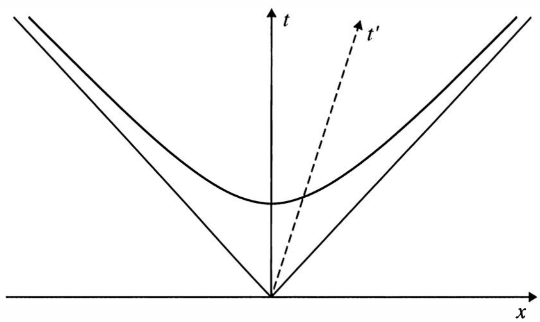


Fig. 2.2 A reconstruction of Minkowski's 5 November, 1907 presentation of relativistic velocity space, with a pair of temporal axes, one spatial axis, a unit hyperbola and its asymptotes

When faced with the question of how to define a four-velocity vector, Minkowski had the option of adopting the definition given by Poincaré in 1905. Instead, he rederived his own version, by following a simple rule. Minkowski defined a four-vector potential, four-current density, and four-force density, all by simply generalizing ordinary three-component vectors to their four-component counterparts. When he came to define four-velocity, he took over the components of the ordinary velocity vector $\mathfrak{w}$ for the spatial part of four-velocity and added an imaginary fourth component, $\mathrm{i}\sqrt{1-\mathfrak{w}^2}$. This resulted in four components of four-velocity, w_1, w_2, w_3, w_4

$$\mathfrak{w}_x\,,\quad \mathfrak{w}_y\,,\quad \mathfrak{w}_z\,,\quad \mathrm{i}\sqrt{1-\mathfrak{w}^2}\,. \tag{2.8}$$

Since the components of Minkowski's quadruplet do not transform like the coordinates of his vector space x_1, x_2, x_3, x_4, they lack what he knew to be a four-vector property.

Minkowski's error in defining four-velocity indicates that he did not yet grasp the notion of four-velocity as a four-vector tangent to the world line of a particle [2.8]. If we grant ourselves the latter notion, then we can let the square of the differential parameter $\mathrm{d}\tau$ of a given world line be $\mathrm{d}\tau^2 = -(\mathrm{d}x_1^2 + \mathrm{d}x_2^2 + \mathrm{d}x_3^2 + \mathrm{d}x_4^2)$, such that the four-velocity w_μ may be defined as the first derivative with respect to τ, $w_\mu = \mathrm{d}x_\mu/\mathrm{d}\tau$ ($\mu = 1, 2, 3, 4$). In addition to a valid four-velocity vector, Minkowski was missing a four-force vector, and a notion of proper time. In light of these significant lacunæ in his knowledge of the basic mathematical objects of four-dimensional physics, *Minkowski*'s triumphant description of his four-dimensional formalism as *virtually the greatest triumph ever shown by the application of mathematics* [2.28, p. 373] is all the more remarkable, and bears witness to the depth of Minkowski's conviction that he was on the right track.

Sometime after Minkowski spoke to the Göttingen Mathematical Society, he repaired his definition of four-velocity, and perhaps in connection with this, he came up with the constitutive elements of his theory of spacetime. In particular, he formulated the idea

of proper time as the parameter of a hyperline in spacetime, the light-hypercone structure of spacetime, and the spacetime equations of motion of a material particle. He expressed his new theory in a 60-page memoir [2.33] published in the *Göttinger Nachrichten* on 5 April, 1908.

His memoir, entitled *The basic equations for electromagnetic processes in moving bodies* made for challenging reading. It was packed with new notation, terminology, and calculation rules, it made scant reference to the scientific literature, and offered no figures or diagrams. Minkowski defined a single differential operator, named *lor* in honor of Lorentz, which streamlined his expressions, while rendering them all the more unfamiliar to physicists used to the three-dimensional operators of ordinary vector analysis.

Along the same lines, Minkowski rewrote velocity, denoted q, in terms of the tangent of an imaginary angle $\mathrm{i}\psi$

$$q = -\mathrm{i} \tan \mathrm{i}\psi \,, \tag{2.9}$$

where $q < 1$. From his earlier geometric interpretation of hyperbolic velocity space, Minkowski kept the idea that every rotation of a t-diameter corresponds to a Lorentz transformation, which he now expressed in terms of $\mathrm{i}\psi$

$$\begin{aligned} x_1' &= x_1\,, \quad x_3' = x_3 \cos \mathrm{i}\psi + x_4 \sin \mathrm{i}\psi\,, \\ x_2' &= x_2\,, \quad x_4' = -x_3 \sin \mathrm{i}\psi + x_4 \cos \mathrm{i}\psi\,. \end{aligned} \tag{2.10}$$

Minkowski was undoubtedly aware of the connection between the composition of Lorentz transformations and velocity composition, but he did not mention it. In fact, Minkowski neither mentioned Einstein's law of velocity addition, nor expressed it mathematically.

While Minkowski made no appeal in *The basic equations* to the hyperbolic geometry of velocity vectors, he retained the hypersurface (2.7) on which it was based and provided a new interpretation of its physical significance. This interpretation represents an important clue to understanding how Minkowski discovered the world line structure of spacetime. The appendix to *The basic equations* rehearses the argument according to which one may choose any point on (2.7) such that the line from this point to the origin forms a new time axis, and corresponds to a Lorentz transformation. He further defined a *spacetime line* to be the totality of spacetime points corresponding to any particular point of matter for all time t.

With respect to the new concept of a spacetime line, Minkowski noted that its direction is determined at every spacetime point. Here Minkowski introduced the notion of *proper time* (*Eigenzeit*), τ, expressing the increase of coordinate time $\mathrm{d}t$ for a point of matter with respect to $\mathrm{d}\tau$

$$\begin{aligned} \mathrm{d}\tau &= \sqrt{\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2} = \mathrm{d}t\sqrt{1-\mathfrak{w}^2} \\ &= \frac{\mathrm{d}x_4}{w_4}\,, \end{aligned} \tag{2.11}$$

where $\mathfrak{w}^2$ is the square of ordinary velocity, $\mathrm{d}x_4 = \mathrm{i}\,\mathrm{d}t$, and $w_4 = \mathrm{i}/\sqrt{1-\mathfrak{w}^2}$, which silently corrects the flawed definition of this fourth component of four-velocity (2.8) delivered by Minkowski in his November 5 lecture.

Although Minkowski did not connect four-velocity to Einstein's law of velocity addition, others did this for him, beginning with *Sommerfeld*, who expressed parallel velocity addition as the sum of tangents of an imaginary angle [2.34]. Minkowski's former student *Philipp Frank* reexpressed both velocity and the Lorentz transformation as hyperbolic functions of a real angle [2.35]. The Serbian mathematician *Vladimir Varičak* found relativity theory to be ripe for application of hyperbolic geometry, and recapitulated several relativistic formulæ in terms of hyperbolic functions of a real angle [2.36]. A small group of mathematicians and physicists pursued this *non-Euclidean style* of Minkowskian relativity, including Varičak, Alfred Robb, Émile Borel, Gilbert Newton Lewis, and Edwin Bidwell Wilson [2.37].

The definition of four-velocity was formally linked by Minkowski to the hyperbolic space of velocity vectors in *The basic equations*, and thereby to the light-cone structure of spacetime. Some time before Minkowski came to study the Lorentz transformation in earnest, both Einstein and Poincaré understood light waves in empty space to be the only physical objects immune to Lorentz contraction. Minkowski noticed that when light rays are considered as world lines, they divide spacetime into three regions, corresponding to the spacetime region inside a future-directed ($t > 0$) hypercone (*Nachkegel*), the region inside a past-directed ($t < 0$) hypercone (*Vorkegel*), and the region outside any such hypercone pair. The propagation in space and time of a spherical light wave is described by a hypercone, or what Minkowski called a light cone (*Lichtkegel*).

One immediate consequence for Minkowski of the light-cone structure of spacetime concerned the relativity of simultaneity. In a section of *The basic equations* entitled *The concept of time*, *Minkowski* [2.33, § 6] showed that Einstein's relativity of simultaneity is not absolute. While the relativity of simultaneity is indeed valid for two or three simultaneous *events* (*Ereignisse*), the simultaneity of four events is absolute, so long as the four spacetime points do not lie on the same spatial plane. Minkowski's demonstration relied on the Einstein simultaneity convention, and employed both light signals and spacetime geometry. His result showed the advantage of employing his spacetime geometry in physics, and later writers – including Poincaré – appear to have agreed with him, by attributing to spacetime geometry the discovery of the existence of a class of events for a given observer that can be the cause of *no other events* for the same observer [2.12, p. 210].

Another signature result of Minkowski's spacetime geometry was the geometric derivation of a Lorentz-covariant law of gravitation. Like Poincaré, Minkowski proposed two four-vector laws of gravitation, exploiting analogies to Newtonian gravitation and Maxwellian electrodynamics, respectively. Minkowski presented only the Newtonian version of the law of gravitation in *The basic equations*, relating the states of two massive particles in arbitrary motion and obtaining an expression for the spacelike component of the four-force of gravitation. Although his derivation involved a new spacetime geometry, Minkowski did not illustrate his new law graphically, a decision which led some physicists to describe his theory as unintelligible. According to Minkowski, however, his achievement was a formal one, inasmuch as Poincaré had formulated his theory of gravitation by proceeding in what he described as a *completely different way* [2.8, p. 225].

Few were impressed at first by Minkowski's innovations in spacetime geometry and four-dimensional vector calculus. Shortly after *The basic equations* appeared in print, two of Minkowski's former students, Einstein and Laub, discovered what they believed to be an infelicity in Minkowski's definition of ponderomotive force density [2.38]. These two young physicists were more impressed by Minkowski's electrodynamics of moving media than by the novel four-dimensional formalism in which it was couched, which seemed far too laborious. Ostensibly as a service to the community, Einstein and Laub reexpressed Minkowski's theory in terms of ordinary vector analysis [2.39, Doc. 51].

Minkowski's reaction to the latter work is unknown, but it must have come to him as a disappointment. According to *Max Born*, Minkowski always aspired [2.40]:

> ... *to find the form for the presentation of his thoughts that corresponded best to the subject matter.*

The form Minkowski gave to his theory of moving media in *The basic equations* had been judged unwieldy by a founder of relativity theory, and in the circumstances, decisive action was called for if his formalism was not to be ignored. In September 1908, during the annual meeting of the German Association of Scientists and Physicians in Cologne, Minkowski took action, by affirming the reality of the four-dimensional *world* and its necessity for physics [2.41]. The next section focuses on the use to which Minkowski put spacetime diagrams in his Cologne lecture, and how these diagrams relate to Poincaré's light ellipse.

2.3 Spacetime Diagrams

One way for Minkowski to persuade physicists of the value of his spacetime approach to understanding physical interactions was to appeal to their visual intuition [2.30]. From the standpoint of visual aids, the contrast between Minkowski's two publications on spacetime is remarkable: where *The basic equations* is bereft of diagrams and illustrations, Minkowski's Cologne lecture makes effective use of diagrams in two and three dimensions. For instance, Minkowski employed two-dimensional spacetime diagrams to illustrate FitzGerald–Lorentz contraction of an electron and the light-cone structure of spacetime (Fig. 2.3).

Minkowski's lecture in Cologne, entitled *Space and time*, offered two diagrammatic readings of the Lorentz transformation, one of his own creation, the other he attributed to Lorentz and Einstein. One of these readings was supposed to represent the kinematics of the theory of relativity of Lorentz and Einstein. In fact, Minkowski's reading captured Lorentzian kinematics, but distorted those of Einstein, and prompted corrective action from Philipp Frank, Guido Castelnuovo, and Max Born [2.42]. The idea stressed by Minkowski was that in the (Galilean) kinematics employed in Lorentz's electron theory, time being absolute, the temporal axis on a space-time diagram may be rotated freely about

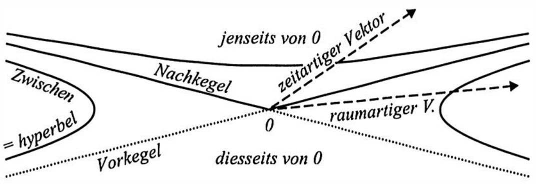


Fig. 2.3 The light-cone structure of spacetime (after [2.1])

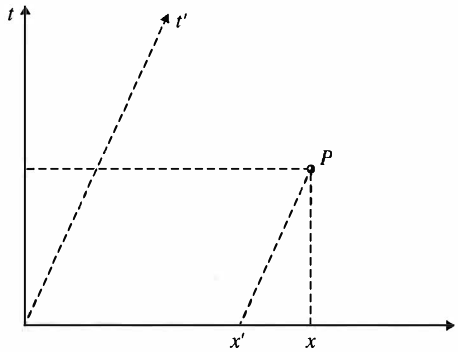


Fig. 2.4 A reconstruction of Minkowski's depiction of the kinematics of Lorentz and Einstein

the coordinate origin in the upper half-plane ($t > 0$), as shown in Fig. 2.4. The location of a point P may be described with respect to frames S and S', corresponding to axes (x, t) and (x', t'), respectively, according to the transformation: $x' = x - vt$, $t' = t$.

In contradistinction to the latter view, the theory proposed by Minkowski required a certain symmetry between the spatial and temporal axes. This constraint on symmetry sufficed for a geometric derivation of the Lorentz transformation. Minkowski described his spacetime diagram (Fig. 2.5) as an illustration of the Lorentz transformation, and provided an idea of a demonstration in *Space and time.* A demonstration was later supplied by Sommerfeld, in an editorial note to his friend's lecture [2.43, p. 37], which appeared in an anthology of papers on the theory of relativity edited by *Otto Blumenthal* [2.44].

Minkowski's spacetime map was not the only illustration of relativistic kinematics available to scientists in the first decade of the twentieth century. Theorists pursuing the non-Euclidean style of Minkowskian relativity had recourse to models of hyperbolic geometry on occasion. The Poincaré half-plane and disk models of hyperbolic geometry were favored by Varičak in this context, for example. Poincaré himself did not employ

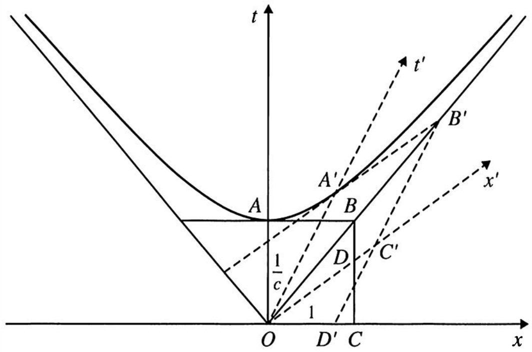


Fig. 2.5 *Minkowski*'s spacetime diagram (after [2.1])

such models in his investigations of the principle of relativity, preferring his light ellipse.

Of these three types of diagram, the light ellipse, spacetime map, and hyperbolic map, only the spacetime map attracted a significant scientific following. The relation between the spacetime map and the hyperbolic maps was underlined by Minkowski, as shown above in relation to surfaces (2.6) and (2.7). There is also a relation between the light ellipse and the spacetime map, although this may not have been apparent to either Poincaré or Minkowski. Their published appreciations of each other's contributions to relativity field the barest of acknowledgments, suggesting no substantial intellectual indebtedness on either side.

The diagrams employed in the field of relativity by Poincaré and Minkowski differ in several respects, but one difference in particular stands out. On the one hand, the light ellipse represents spatial relations in a plane defined as a meridional section of an ellipsoid of rotation. A Minkowski diagram, on the other hand, involves a temporal axis in addition to a spatial axis (or two, for a three-dimensional spacetime map). This difference does not preclude representation of a light ellipse on a Minkowski diagram, as shown in Figs. 2.6 and 2.7, corresponding, respectively, to the two interpretations of the Lorentz transformation offered by Poincaré before and after 1909.

In Poincaré's pre-1909 interpretation of the Lorentz transformation, the radius vector of the light ellipse corresponds to light points at an instant of time t as read by clocks at rest in the ether frame. The representation of this situation on a Minkowski diagram is that of an ellipse contained in a spacelike plane of constant time t (Fig. 2.6). The ellipse center coincides with spacetime point $B = (vt, 0, t)$, and the points E, B, F, and A lie on the major axis, such that BH is a semi-minor axis of

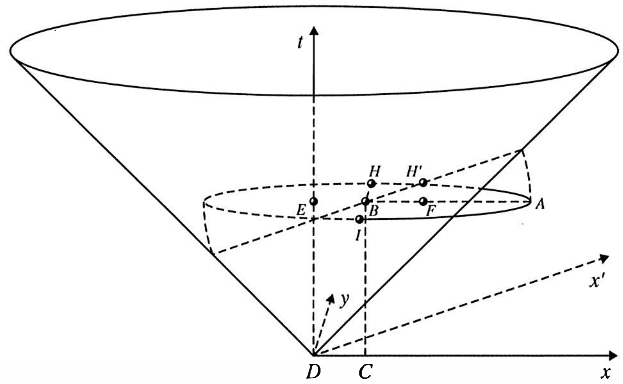


Fig. 2.6 Spacetime model of Poincaré's light ellipse (1906) in a spacelike plane ($t = \text{const.}$)

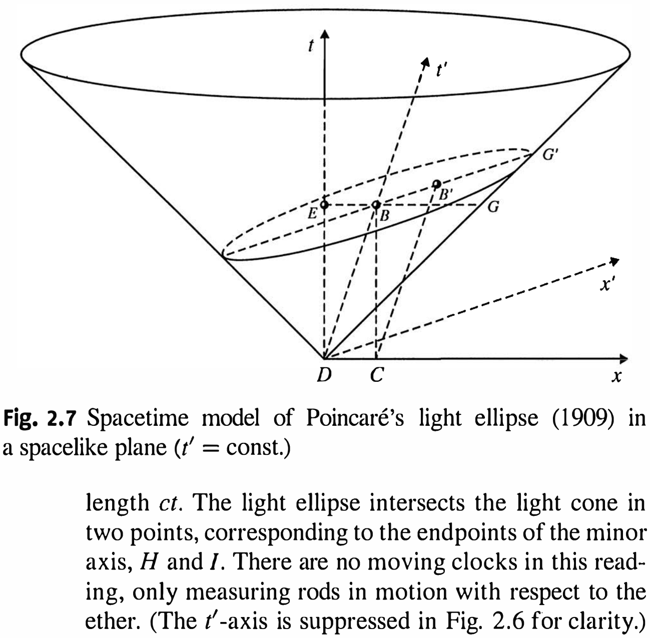


Fig. 2.7 Spacetime model of Poincaré's light ellipse (1909) in a spacelike plane ($t' = \text{const.}$)

length ct. The light ellipse intersects the light cone in two points, corresponding to the endpoints of the minor axis, H and I. There are no moving clocks in this reading, only measuring rods in motion with respect to the ether. (The t'-axis is suppressed in Fig. 2.6 for clarity.) The abstract nature of Poincaré's early interpretation of the light ellipse is apparent in the Minkowskian representation, in that there are points on the light ellipse that lie outside the light cone, and are physically inaccessible to an observer at rest in the ether.

In Poincaré's post-1909 repurposing of the light ellipse, the light pulse is measured with comoving clocks, such that the corresponding figure on a Minkowski diagram is an ellipse in a plane of constant time t'. The latter $x'y'$-plane intersects the light cone at an oblique angle, as shown in Fig. 2.7, such that their intersection is a Poincaré light ellipse. (The y-axis and the y'-axis are suppressed for clarity.)

Both before and after 1909, Poincaré found that a spherical light pulse in the ether would be described as a prolate ellipsoid in inertial frames. Meanwhile, for Einstein and others who admitted the spatio-temporal relativity of inertial frames, the form of a spherical light pulse remained spherical in all inertial frames. In Poincaré's scheme of things, the light pulse is a sphere only for ether-fixed observers measuring wavefronts with clocks and rods at rest; in all other inertial frames the light pulse is necessarily shaped like a prolate ellipsoid.

Comparison of Poincaré's pre-1909 and post-1909 readings of the light ellipse shows the ellipse dimensions to be unchanged. What differs in the Minkowskian representations of these two readings is the angle of the spacelike plane containing the light ellipse with respect to the light cone. The complementary representation is obtained in either case by rotating the light ellipse through an angle $\psi = \tanh^{-1} v$ about the line parallel to the y-axis passing through point B.

We are now in a position to answer the question raised above, concerning the reasons for Poincaré's embrace of time deformation in 1909. From the standpoint of experiment, there was no pressing need to recognize time deformation in 1909, although in 1907 *Einstein* figured it would be seen as a transverse Doppler effect in the spectrum of canal rays [2.45]. On the theoretical side, Minkowski's spacetime theory was instrumental in convincing leading ether-theorists like Sommerfeld and Max Abraham of the advantages of Einstein's theory. Taken in historical context, Poincaré's poignant acknowledgment in Göttingen of time deformation (and subsequent repurposing of his light ellipse) reflects the growing appreciation among scientists, circa 1909, of the Einstein–Minkowski theory of relativity [2.46].

References

2.1 H. Minkowski: Raum und Zeit, Jahresber. Dtsch. Math.-Ver. **18**, 75–88 (1909)
2.2 H.A. Lorentz: Electromagnetic phenomena in a system moving with any velocity less than that of light, Proc. Sect. Sci. K. Akad. Wet. Amst. **6**, 809–831 (1904)
2.3 S. Walter, E. Bolmont, A. Coret (Eds.): *La Correspondance d'Henri Poincaré, Vol. 2: La correspondance entre Henri Poincaré et les physiciens, chimistes et ingénieurs* (Birkhäuser, Basel 2007)
2.4 H. Poincaré: Sur la dynamique de l'électron, Rend. Circ. Mat. Palermo **21**, 129–176 (1906)
2.5 A. Einstein: Zur Elektrodynamik bewegter Körper, Ann. Phys. **17**, 891–921 (1905)
2.6 J. Gray, S. Walter (Eds.): *Henri Poincaré: Trois suppléments sur la découverte des fonctions fuchsiennes* (Akademie, Berlin 1997)
2.7 S. Lie, G. Scheffers: *Vorlesungen über continuierliche Gruppen mit geometrischen und anderen Anwendungen* (Teubner, Leipzig 1893)
2.8 S. Walter: Breaking in the 4-vectors: The four-dimensional movement in gravitation, 1905–1910. In: *The Genesis of General Relativity*, Vol. 3, ed. by J. Renn, M. Schemmel (Springer, Berlin 2007) pp. 193–252
2.9 R. Marcolongo: Sugli integrali delle equazioni dell' elettrodinamica, Atti della R. Accademia dei Lincei, Rend. Cl. Sci. Fis. Mat. Nat. **15**, 344–349 (1906)
2.10 M.J. Crowe: *A History of Vector Analysis: The Evolution of the Idea of a Vectorial System* (Univ. Notre Dame Press, South Bend 1967)
2.11 H. Poincaré: *The Value of Science: Essential Writings of Henri Poincaré* (Random House, New York 2001)
2.12 S. Walter: Hypothesis and convention in Poincaré's defense of Galilei spacetime. In: *The Significance of the Hypothetical in the Natural Sciences*, ed. by M. Heidelberger, G. Schiemann (de Gruyter, Berlin 2009) pp. 193–219
2.13 H. Poincaré: L'avenir des mathématiques, Rev. Gén. Sci. Pures Appl. **19**, 930–939 (1908)
2.14 O. Darrigol: Poincaré and light, Poincaré, 1912–2012, Séminaire Poincaré 16 (École polytechnique, Palaiseau 2012) pp. 1–43
2.15 H. Poincaré: *Sechs Vorträge über ausgewählte Gegenstände aus der reinen Mathematik und mathematischen Physik* (Teubner, Leipzig Berlin 1910)
2.16 H. Poincaré: La mécanique nouvelle, Rev. Sci. **12**, 170–177 (1909)
2.17 A.I. Miller: *Albert Einstein's Special Theory of Relativity: Emergence (1905) and Early Interpretation* (Addison-Wesley, Reading, MA 1981)
2.18 J. Schwermer: Räumliche Anschauung und Minima positiv definiter quadratischer Formen, Jahresber. Dtsch. Math.-Ver. **93**, 49–105 (1991)
2.19 D.E. Rowe: 'Jewish mathematics' at Göttingen in the era of Felix Klein, Isis **77**, 422–449 (1986)
2.20 H. Minkowski: *Geometrie der Zahlen* (Teubner, Leipzig 1896)
2.21 L. Corry: *David Hilbert and the Axiomatization of Physics (1898–1918): From Grundlagen der Geometrie to Grundlagen der Physik* (Kluwer, Dordrecht 2004)
2.22 D. Cahan: The institutional revolution in German physics, 1865-1914, Hist. Stud. Phys. Sci. **15**, 1–65 (1985)
2.23 C. Jungnickel, R. McCormmach: *Intellectual Mastery of Nature: Theoretical Physics from Ohm to Einstein* (University of Chicago Press, Chicago 1986)
2.24 L. Pyenson: Physics in the shadow of mathematics: the Göttingen electron-theory seminar of 1905, Arch. Hist. Exact Sci. **21**(1), 55–89 (1979)
2.25 S. Walter: Hermann Minkowski's approach to physics, Math. Semesterber. **55**(2), 213–235 (2008)
2.26 M. Planck: Zur Dynamik bewegter Systeme, Sitzungsber. k. preuss. Akad. Wiss. 542–570 (1907)
2.27 M.J. Klein, A.J. Kox, R. Schulmann (Eds.): *The Collected Papers of Albert Einstein, Vol. 5, The Swiss Years: Correspondence, 1902–1914* (Princeton University Press, Princeton 1993)
2.28 H. Minkowski: Das Relativitätsprinzip, Jahresber. Dtsch. Math.-Ver. **24**, 372–382 (1915)
2.29 S. Walter: La vérité en géométrie: sur le rejet mathématique de la doctrine conventionnaliste, Philos. Sci. **2**, 103–135 (1997)
2.30 P. Galison: Minkowski's spacetime: from visual thinking to the absolute world, Hist. Stud. Phys. Sci. **10**, 85–121 (1979)
2.31 H. von Helmholtz: *Vorträge und Reden*, 3rd edn. (Vieweg, Braunschweig 1884)
2.32 W.F. Reynolds: Hyperbolic geometry on a hyperboloid, Am. Math. Mon. **100**, 442–455 (1993)
2.33 H. Minkowski: Die Grundgleichungen für die elektromagnetischen Vorgänge in bewegten Körpern, Nachr. K. Ges. Wiss. Göttingen, 53–111 (1908)
2.34 A. Sommerfeld: Über die Zusammensetzung der Geschwindigkeiten in der Relativtheorie, Phys. Z. **10**, 826–829 (1909)
2.35 P.G. Frank: Die Stellung des Relativitätsprinzips im System der Mechanik und der Elektrodynamik, Sitzungsber. Kais. Akad. Wiss. Wien IIA **118**, 373–446 (1909)
2.36 V. Varičak: Anwendung der Lobatschefskijschen Geometrie in der Relativtheorie, Phys. Z. **11**, 93–96 (1910)
2.37 S. Walter: The non-Euclidean style of Minkowskian relativity. In: *The Symbolic Universe: Geometry and Physics, 1890–1930*, ed. by J. Gray (Oxford University Press, Oxford 1999) pp. 91–127
2.38 J.J. Stachel, D.C. Cassidy, J. Renn, R. Schulmann: Einstein and Laub on the electrodynamics of moving media. In: *The Collected Papers of Albert Einstein, Vol. 2, The Swiss Years: Writings, 1900–1909*, ed. by J.J. Stachel, D.C. Cassidy, J. Renn, R. Schulmann

(Princeton University Press, Princeton 1989) pp. 503–507
2.39 J.J. Stachel, D.C. Cassidy, J. Renn, R. Schulmann (Eds.): *The Collected Papers of Albert Einstein, Vol. 2, The Swiss Years: Writings, 1900–1909* (Princeton University Press, Princeton 1989)
2.40 M. Born: Besprechung von Max Weinstein, Die Physik der bewegten Materie und die Relativitätstheorie, Phys. Z. **15**, 676 (1914)
2.41 S. Walter: Minkowski's modern world. In: *Minkowski Spacetime: A Hundred Years Later*, ed. by V. Petkov (Springer, Berlin 2010) pp. 43–61
2.42 S. Walter: Minkowski, mathematicians, and the mathematical theory of relativity. In: *The Expanding Worlds of General Relativity*, (Birkhäuser, Boston Basel 1999) pp. 45–86
2.43 D.E. Rowe: A look back at Hermann Minkowski's Cologne lecture 'Raum und Zeit', Math. Intell. **31**(2), 27–39 (2009)
2.44 O. Blumenthal (Ed.): *Das Relativitätsprinzip; Eine Sammlung von Abhandlungen* (Teubner, Leipzig 1913)
2.45 A. Einstein: Über die Möglichkeit einer neuen Prüfung des Relativitätsprinzips, Ann. Phys. **23**, 197–198 (1907)
2.46 S. Walter: Poincaré on clocks in motion, Stud. Hist. Philos. Modern Phys. (2014), doi: 10.1016/j.shpsb.2014.01.003

3. Relativity Today

Nick M. J. Woodhouse

This chapter outlines the special theory of relativity from a modern as opposed to historical perspective. It follows the approach promoted by Hermann Bondi, in which measuring rods and rigid frames of reference take second place to an exploration of the geometry of spacetime through thought experiments involving light signals and clocks in uniform motion. The theory is developed up to the point of demonstrating that Maxwell's equations and the principle of relativity are compatible within the framework of Minkowski space.

Special relativity had a difficult beginning, growing as it did out of the wreckage of the ether theory. The physics community was slow to accept Einstein's radical insight for reasons that are now hard to appreciate. The confused and confusing challenges over the twin *paradox* and other misunderstandings were slow to fade.
Today the ideas are not at the frontier of understanding, thinly supported by a few delicate and subtle experiments, but rather the stuff of engineering: global positioning system (GPS) devices, particle accelerators, and other modern machines, simply would not work if their designs were not based on relativistic calculations.

Nonetheless relativity still challenges intuition. It is not the primacy of the principle of relativity that does that. The idea that the behavior of dynamical systems is the same in all inertial frames of frames of refer-

ence goes back to Newton, and before him to Galileo. Nor is it that special relativity introduces any unfamiliar persona into the play: space, time, mass, momentum, charge and so on all appear in the classical story. The difficulty is in letting go of concepts that seem to be integral to our understanding of the world, notably absolute space and universal time.

An approach to the theory that follows the historical development of the ideas or builds bridges from the classical ideas of space and time is likely to generate the same resistance to the unfamiliar, and the same confusions in which the new worldview is overlayed by an imperfectly excluded classical picture. For this reason, this chapter follows the approach principally promoted by Hermann Bondi, in which measuring rods and rigid frames of reference take second place to an exploration of the geometry of spacetime through thought experiments involving light signals and clocks in uniform motion. It is *modern* in the sense that the development is not *historical*. A more extended version of the material in this chapter is given in [3.1].

The starting point is an extended form of the principle of relativity that Newton would have recognized in the narrower context of classical mechanics:

> *The observed behavior of all physical systems, including electromagnetic fields, is the same for all observers in uniform motion.*

This is distilled from the work that culminated in Einstein's 1905 paper. It hides some assumptions that should be made explicit. First, there is no medium for electromagnetic waves, no ether: such a medium would determine a preferred standard of rest, in the way that the air does for sound waves. Second, the notion of *uniform motion* makes sense. In other words an observer can identify the absence of acceleration. Here the assumption is that there is no gravity, so any apparent gravitational force can be understood simply as the consequence of acceleration. Third, the behavior of clocks is not affected by nonaccelerating motion. So it makes sense to picture each observer in uniform motion as being equipped with a standard clock, which measures time at the observer's location.

We shall build *special relativity* on this principle as a theory of the geometry of spacetime rather than as theory of the transformations between inertial coordinate systems. Coordinates are of course important, and come later. Without them one cannot make quantitative predictions. But the central point is that the intuitive picture that should replace absolute space and universal time should be of a four-dimensional space of events, the *spacetime* of the title, in which inertial coordinates are convenient labels, just as Cartesian coordinates are convenient labels for points rather than intrinsic structures in Euclidean geometry.

It is worth noting two points. First, that a part of the difficulty is in the tendency to cling to intuitive ideas of space that should be abolished even within the theoretical framework that Newton and Galileo knew. If one takes seriously that all inertial frames in classical physics are on an equal footing, then there is no absolute notion of *location*: the statement that two nonsimultaneous events happened in the same place can be true for one observer, but untrue for another. *Colocation* in classical physics is relative. If one understands why that is so, then it is easier to accept, in special relativity, that simultaneity is also relative.

Beyond that, our intuition even for three-dimensional geometry is limited. Astronauts notoriously lose their way in the weightless environment of the International Space Station (ISS) without the familiar prompt of *up* and *down*. It is therefore understandable that building an intuitive picture of the four-dimensional spacetime in which we live is challenging.

Second, much of this handbook is concerned with gravity, in which context our assumptions are not valid. So it is tempting to regard *special* relativity as an obsolete theory, to be superseded by the general theory, and to take the view that in developing the modern view of fundamental physics, we should first understand the general theory, and then let *special* relativity emerge as an approximation, applicable in a limiting sense when the effects of gravity are negligible, but ultimately misleading as a picture of the world.

There is some force in this, but the counter is that special relativity is incorporated into general relativity not only in this limiting sense, but also in another, as the model of spacetime experienced by an observer in free fall, over short times and distances.

3.1 Operational Definitions

The task is to build a picture of spacetime from the principle of relativity without importing ideas about the measurement of space and time based on limited classical intuition. It requires us to go back to first prin-

ciples, beginning with the measurement of distance: *distance* cannot be taken as a self-evident concept, which requires no closer examination. It must be given an *operational definition* – a definition in terms of the operations required to measure it.

Before tackling distance, however, it is necessary first to give an operational definition of *simultaneity*, because even in the classical view of space and time, distance is defined independently of the motion of the observer only between simultaneous events. Here we use definitions due to *Milne* [3.2] and *Bondi* [3.3].

In Milne's approach, one takes *clocks* and *light signals* as fundamental. Every observer carries a clock with which he can measure the time of events in his immediate vicinity and observers can send out and receive light signals, which are carried by photons (particles of light).

Because Maxwell's equations are to hold in every frame, the definitions of distance and simultaneity must be such that the following is true:

> *The velocity of photons is the same irrespective of the motion of their source or of the observer.*

A nonaccelerating observer moving along a straight line can use his clock and light signals to assign coordinates t and x to distant events on the line, as follows. Suppose that he sends out a light signal at time t_1 (measured on his clock). This is received at an event A on the line and immediately transmitted back to the observer, arriving at time t_2 (again measured on the observer's clock).

If the velocity of photons is assumed to be constant, then the journeys of the outgoing and returning photons will be reckoned by the observer to have equal duration, and so the observer will take the event B that happens at his location at time $\frac{1}{2}(t_1+t_2)$ to be simultaneous with A and he will assign this value of t to A. This is the *radar definition* of simultaneity. It is illustrated in the spacetime diagram (Fig. 3.1), in which each point represents an event and time increases up the page. The vertical line is the history or *worldline* of the observer and the lines at 45° are the worldlines of the outgoing and returning photons.

> *The observer defines A to be simultaneous with the event B on his worldline that happens at time $\frac{1}{2}(t_1+t_2)$ and assigns a distance $\frac{1}{2}c(t_2-t_1)$ to the separation of B from A.*

Here c is a constant that is chosen arbitrarily according to the system units employed, but is given the same value by all observers. By defining distance and simultaneity in this way, a nonaccelerating observer can, in principle, set up a coordinate system to label each event by its radar distance x from his own location, and the time t at which it happens, according to the radar definition. These labels are the *inertial coordinates* of special relativity.

3.1.1 Relativity of Simultaneity

With these definitions, the velocity of light is independent of the observer, but simultaneity is relative. Two events that are reckoned to be simultaneous by one observer O may not be simultaneous according to a second observer O' moving relative to O. If O sets up the inertial coordinate system x, t so that O' passes O at $t=0$, then the worldline of O' is given by $x=ut$ for some constant u, which O will interpret as the velocity of O' (Fig. 3.2).

Consider the event A with coordinates

$$t=0\,,\quad x=-D\quad (D>0)\,.$$

A photon that reaches $x=-D$ at $t=0$ must have left O at time $t=-D/c$ (measured on the clock carried by O).

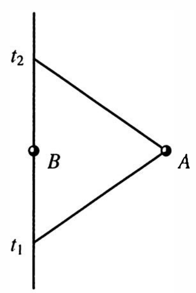


Fig. 3.1 Milne's definitions

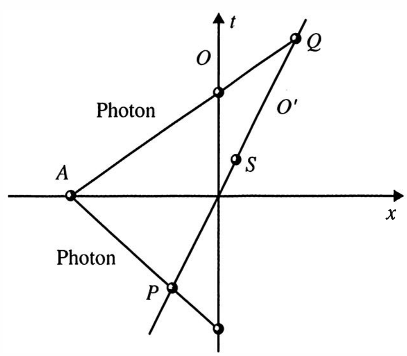


Fig. 3.2 Simultaneity

This photon passes O' at the event P with coordinates x_P, t_P, where $x_P = ut_P = -D - ct_P$. Thus P has coordinates

$$x_P = -\frac{uD}{u+c}, \quad t_P = -\frac{D}{u+c}.$$

Similarly, a photon emitted at A and traveling with the speed c in the positive x-direction reaches O at the event with coordinates $(x,t) = (0, D/c)$ and reaches O' at the event Q with coordinates

$$x_Q = \frac{uD}{c-u}, \quad t_Q = \frac{D}{c-u}.$$

As one would expect, O thinks that A is simultaneous with the origin $(0,0)$ of the inertial coordinates, since this is the event at his own location that happens at time

$$t = \tfrac{1}{2}\left(\frac{D}{c} - \frac{D}{c}\right) = 0.$$

On the other hand, O' thinks that A happens simultaneously with the event at *his* location which happens midway between P and Q. That is, he reckons A is simultaneous with the event S with coordinates

$$x_S = \tfrac{1}{2}(x_P + x_Q) = \frac{u^2 D}{c^2 - u^2},$$
$$t_S = \tfrac{1}{2}(t_P + t_Q) = \frac{uD}{c^2 - u^2}.$$

Since $x_S = ut_S$, this event is indeed on the worldline of O'. However, t_S is nonzero unless either D or u vanishes, so S is not simultaneous with A according to O. Our two observers O and O' have different notions of simultaneity.

If D is 10 m and u is 10 m/s, then with $c = 3 \times 10^8$, we get $t_S \approx 10^{-15}$ s, that is, a femtosecond. To get something more easily observable, either u must be a substantial fraction of the velocity of light or D must be large. If, for example, D is 10 million light years and u is again 10 m/s, then t_S is about 4 months. So even if the two observers have a relative speed as low as 10 m/s, their notions of simultaneity over intergalactic distances are significantly different.

3.1.2 Bondi's *k*-Factor

Consider two observers O and O' traveling along the line with constant speeds (Fig. 3.3). Also suppose that they pass each other at the event E and then move directly away from each other. Suppose also that they both set their clocks to zero at E.

By using the radar definitions, they both set up inertial coordinate systems on two-dimensional spacetime: O will label the events on the line by their distance x along the line, and by the time at which they happen, according to his measurements; and likewise O' will label them by x' and t'.

We shall derive the relationship between (x,t) and (x',t') by making two assumptions, that both observers reckon that the velocity of light is c, and that only their relative motion is observable.

We begin by considering a photon emitted by O toward O' at time t (measured on the clock that O carries). Suppose that it is received by O' at time $t' = kt$ (measured on the clock carried by O'). The quantity k is called *Bondi's k-factor*. Since neither observer is accelerating, k is constant and, as a consequence of the second assumption, *k depends only on the relative velocity of O and O'*. It is in this last innocuous looking statement that we depart from classical ideas.

3.1.3 Time Dilation

Because k depends only on the relative motion, we have the following:

- A photon sent by O toward O' at time t (measured on the clock carried by O) arrives at O' at time $t' = kt$ (measured on the clock carried by O').
- A photon sent by O' toward O at time t' (measured on the clock carried by O') arrives at O at time $t = kt'$ (measured on the clock carried by O).

Now consider the spacetime diagram in Fig. 3.3.

Here a photon sent by O at time t measured on the clock carried by O arrives at O' at event B, which happens at time kt measured on the clock carried by O'; it is then sent back to O, arriving at time $k^2 t$, as measured on

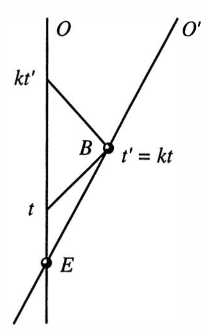


Fig. 3.3 The k-factor

the clock carried by O. Hence O measures the distance from his location to B and the time of B to be

$$d_B = \tfrac{1}{2}c(k^2-1)t\,, \quad t_B = \tfrac{1}{2}(1+k^2)t\,.$$

Thus O reckons that the speed of O' is

$$u = \frac{d_B}{t_B} = \frac{c(k^2-1)}{k^2+1}\,.$$

Solving for k, we have

$$k = \sqrt{\frac{c+u}{c-u}} > 1\,.$$

It follows that

$$\frac{\text{time } E \text{ to } B \text{ measured by } O}{\text{time } E \text{ to } B \text{ measured by } O'} = \frac{t_B}{kt} = \frac{(k^2+1)t}{2kt} = \gamma(u)\,,$$

where the *gamma factor* $\gamma(u)$ is defined by

$$\gamma(u) = \frac{1}{\sqrt{1-\frac{u^2}{c^2}}}\,.$$

This is the *time dilation effect*; the time between the two events depends on the observer. It is paradoxical only if one insists on thinking about *time* independently of the process of measurement of time.

3.2 Lorentz Transformations in Two Dimensions

For simplicity, we shall assume that both observers set their clocks to zero at the event E at which they pass. Then E will be the common origin of the two coordinate systems.

Proposition 3.1
The inertial coordinate systems set up by O and O' are related by

$$\begin{pmatrix} ct \\ x \end{pmatrix} = \gamma(u)\begin{pmatrix} 1 & \frac{u}{c} \\ \frac{u}{c} & 1 \end{pmatrix}\begin{pmatrix} ct' \\ x' \end{pmatrix}\,, \tag{3.1}$$

where u is the relative velocity and $\gamma(u) = 1/\sqrt{1-u^2/c^2}$.

This is the (two-dimensional) *Lorentz transformation.*

Proof: Let k denote Bondi's factor. Consider the spacetime diagram in Fig. 3.4. A photon is sent out from O at time T measured on the clock carried by O, passes O' at time kT measured on the clock carried by O', is reflected at the event B, passes O' again at time T' measured on the clock carried by O', and returns to O at time kT' measured on the clock carried by O.

In the inertial coordinate system of observer O, the coordinates of B are

$$t = \tfrac{1}{2}(kT'+T)\,, \quad x = \tfrac{1}{2}c(kT'-T)\,.$$

While in the inertial coordinate system of observer O', the coordinates of B are

$$t' = \tfrac{1}{2}(T'+kT)\,, \quad x' = \tfrac{1}{2}c(T'-kT)\,.$$

Hence we have

$$\begin{pmatrix} ct \\ x \end{pmatrix} = \frac{c}{2}\begin{pmatrix} 1 & k \\ -1 & k \end{pmatrix}\begin{pmatrix} T \\ T' \end{pmatrix}\,,$$

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \frac{c}{2}\begin{pmatrix} k & 1 \\ -k & 1 \end{pmatrix}\begin{pmatrix} T \\ T' \end{pmatrix}\,,$$

and therefore

$$\begin{pmatrix} ct \\ x \end{pmatrix} = \frac{1}{2}\begin{pmatrix} k+k^{-1} & k-k^{-1} \\ k-k^{-1} & k+k^{-1} \end{pmatrix}\begin{pmatrix} ct' \\ x' \end{pmatrix}.$$

But we showed above that $k = \sqrt{(c+u)/(c-u)}$. Therefore,

$$k+k^{-1} = \frac{2c}{\sqrt{c^2-u^2}}\,, \quad k-k^{-1} = \frac{2u}{\sqrt{c^2-u^2}}\,,$$

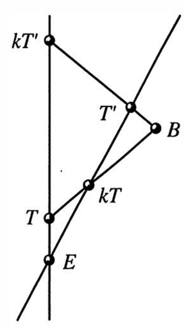


Fig. 3.4 The derivation of the coordinate transformation

and the result follows. Some sign choices have been made in relating x and x', which can be positive or negative, to the distances from O or O', which are necessarily positive. ■

The relationship between the two coordinate systems is shown in Fig. 3.5. If we put $x' = 0$, then $x = ut$.

With the sign choices we have made, O' is moving relative to O in the positive x direction with speed u. We also have $t = \gamma(u)t'$ when $x' = 0$, which is the time dilation formula for events on the worldline of O'. The inverse transformation is

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \gamma(u) \begin{pmatrix} 1 & -\frac{u}{c} \\ -\frac{u}{c} & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix},$$

so O is moving relative to O' in the *negative* x' direction with the same speed u.

The transformation reduces to the Galilean transformation when $u \ll c$, since we have

$$\begin{pmatrix} t \\ x \end{pmatrix} = \gamma(u) \begin{pmatrix} 1 & \frac{u}{c^2} \\ u & 1 \end{pmatrix} \begin{pmatrix} t' \\ x' \end{pmatrix} \to \begin{pmatrix} 1 & 0 \\ u & 1 \end{pmatrix} \begin{pmatrix} t' \\ x' \end{pmatrix}$$

as $c \to \infty$.

3.2.1 Transformation of Velocity

Consider a nonaccelerating particle moving with speed v relative to O in the negative x direction, so that $x = -vt + a$ for some constant a. In classical theory, its speed relative to O' would be $u + v$. In special relativity, we have

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \gamma(u) \begin{pmatrix} 1 & -\frac{u}{c} \\ -\frac{u}{c} & 1 \end{pmatrix} \begin{pmatrix} ct \\ -vt + a \end{pmatrix},$$

so in the x', t' coordinate system set up by O', the events in the history of the particle are given by

$$t' = \gamma(u) \left[\left(1 + \frac{uv}{c^2}\right) t - \frac{au}{c^2} \right]$$
$$x' = \gamma(u) \left[-(u + v)t + a \right].$$

Therefore, the speed w of the particle relative to O' is

$$w = -\frac{dx'}{dt'} = \frac{v + u}{1 + \frac{uv}{c^2}}.$$

This is the *velocity addition formula*. Note that if $|u| < c$ and $|v| < c$ then $|w| < c$ and that the formula reduces to $w = v + u$ when $|u|, |v| \ll c$.

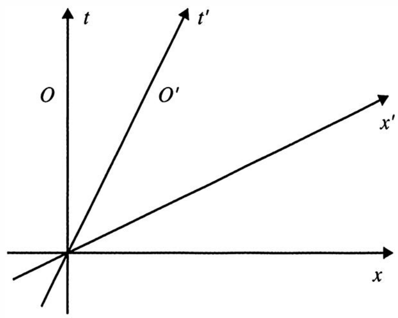


Fig. 3.5 The two-dimensional Lorentz transformation

3.2.2 Lorentz Contraction

Consider two observers O and O' whose inertial coordinate systems are related by (3.1). Suppose that a rod lies along the x'-axis between $x' = 0$ and $x' = L$ and is at rest relative to O'. Then according to O', its length is L. What is its length as measured by O?

We must first be clear about what the question means. In the inertial coordinate system of O', the worldlines of the ends of the rod are given by $x' = 0$ and by $x' = L$. In the inertial coordinate system of O, therefore, the two worldlines are given parametrically (with t' as parameter) by

$$\begin{pmatrix} ct \\ x \end{pmatrix} = \gamma(u) \begin{pmatrix} 1 & \frac{u}{c} \\ \frac{u}{c} & 1 \end{pmatrix} \begin{pmatrix} ct' \\ 0 \end{pmatrix} = \gamma(u) \begin{pmatrix} ct' \\ ut' \end{pmatrix}, \tag{3.2}$$

$$\begin{pmatrix} ct \\ x \end{pmatrix} = \gamma(u) \begin{pmatrix} 1 & \frac{u}{c} \\ \frac{u}{c} & 1 \end{pmatrix} \begin{pmatrix} ct' \\ L \end{pmatrix} = \gamma(u) \begin{pmatrix} ct' + \frac{Lu}{c} \\ ut' + L \end{pmatrix}. \tag{3.3}$$

The question is: What is the distance measured by O between two events E and B, one on the first worldline, one on the second, which are simultaneous according to O? If we take E to be the event $t = 0$, $x = 0$, then B must be as in (3.3), with t' chosen so that $t = 0$. That is $t' = -Lu/c^2$, which implies that B is the event

$$t = 0, \quad x = \gamma(u)\left(-\frac{Lu^2}{c^2} + L\right) = L\sqrt{1 - \frac{u^2}{c^2}}. \tag{3.4}$$

So according to O, the rod is shorter by a factor $\sqrt{1 - u^2/c^2}$.

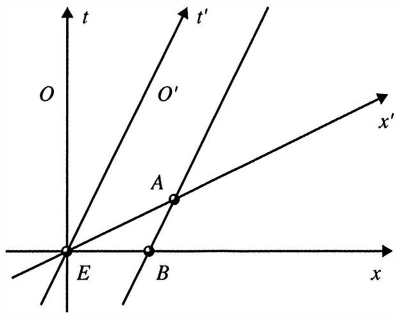


Fig. 3.6 The Lorentz contraction

3.2.3 Composition of Lorentz Transformations

The composition of two Lorentz transformations with velocities u and v is a Lorentz transformation with velocity w, where

$$\gamma(w)\begin{pmatrix}1 & \frac{w}{c}\\ \frac{w}{c} & 1\end{pmatrix} = \gamma(u)\gamma(v)\begin{pmatrix}1 & \frac{v}{c}\\ \frac{v}{c} & 1\end{pmatrix}\begin{pmatrix}1 & \frac{u}{c}\\ \frac{u}{c} & 1\end{pmatrix},$$

with

$$\gamma(w) = \gamma(u)\gamma(v)\left(1+\frac{uv}{c^2}\right) . \tag{3.5}$$

However,

$$\frac{w^2}{c^2} = \frac{\gamma(w)^2 - 1}{\gamma(w)^2} .$$

After substituting from (3.5), and doing a little algebra, we find that

$$w^2 = \frac{(u+v)^2}{\left(1+\frac{uv}{c^2}\right)^2} ,$$

which again gives the velocity addition formula.

3.2.4 Rapidity

The Lorentz transformation and velocity addition formula take on a more familiar look if we put $\phi(u) = \log k = \tanh^{-1}(u/c)$. Then

$$\begin{pmatrix}ct\\ x\end{pmatrix} = \begin{pmatrix}\cosh\phi & \sinh\phi\\ \sinh\phi & \cosh\phi\end{pmatrix}\begin{pmatrix}ct'\\ x'\end{pmatrix} ,$$

so a Lorentz transformation is a *hyperbolic rotation.* The quantity ϕ is called the *rapidity* or *pseudo-velocity* of the transformation. It is analogous to the angle of a rotation in the plane.

In terms of rapidity, the velocity addition formula takes the more suggestive form

$$\phi(w) = \phi(u) + \phi(v) .$$

3.2.5 Lorentz and Poincaré Groups

The Lorentz transformations in two-dimensional spacetime form a group, called the *Lorentz group* (in two dimensions). Rapidity determines an isomorphism with $\mathbb{R}$ (under addition). More precisely, this group is the *proper orthochronous Lorentz group*: the full group is obtained by composing Lorentz transformations with:

- The *space reflection*

$$\begin{pmatrix}ct\\ x\end{pmatrix} \mapsto \begin{pmatrix}ct'\\ x'\end{pmatrix} = \begin{pmatrix}ct\\ -x\end{pmatrix}$$

  which reverse the orientation of the line
- The *time reversal*

$$\begin{pmatrix}ct\\ x\end{pmatrix} \mapsto \begin{pmatrix}ct'\\ x'\end{pmatrix} = \begin{pmatrix}-ct\\ x\end{pmatrix} .$$

This gives a *Lie group* with four connected components, each homeomorphic to the real line. Which component a transformation lies in is determined by whether it reverses time, or spatial orientation, or both, or neither. The analogous group in Euclidean geometry is O(2), but this has only two components.

The full Lorentz group is extended to give the *Poincaré group*, which is generated by (proper, orthochronous) Lorentz transformations, space reflections, time reversals, and translations. In the following, all Lorentz transformations will be proper and orthochronous, unless we explicitly allow otherwise.

3.3 Inertial Coordinates in Four Dimensions

An observer traveling in a straight line at constant speed can determine the coordinates t, x of events that happen along the line by the radar method. In order to assign coordinates to events that happen elsewhere in space, an observer needs, in addition to a clock, a device to measure the direction from which light signals arrive. He can then assign spherical polar coordinates to an event; the distance r from his own location and the time of the event are defined by the radar method; and the two polar angular coordinates θ and ϕ are given by the direction of the returning light signal. From r, θ, ϕ he can recover the Cartesian coordinates x, y, z of the event by using the standard transformation.

The result is an *inertial frame of reference* or *inertial coordinate system* t, x, y, z on spacetime, in which the observer's own worldline is given by

$$x = y = z = 0\,.$$

Implicit in this operational definition is the assumption that the observer knows how to compare the directions from which light signals arrive at different times; in other words it makes sense to say that the angle-measuring device is carried *without rotation.* An *inertial* observer is an observer who is neither accelerating nor rotating.

3.3.1 Four-Dimensional Coordinate Transformations

In order to derive the relationship between the coordinate systems t, x, y, z and t', x', y', z' set up by two inertial observers O and O', we have to make some assumptions.

- The transformation is affine linear. That is, it is of the form

$$\begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} = L \begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} + C\,, \tag{3.6}$$

  where L is a nonsingular 4×4 matrix and C is a column vector.
- Photons travel in straight lines with velocity c relative to any inertial coordinate system. That is, photon worldlines are of the form

$$x = u_1 t + a_1\,, \quad y = u_2 t + a_2\,, \quad z = u_3 t + a_3\,,$$

  where u_1, u_2, u_3, a_1, a_2, a_3 are constants and $u_1^2 + u_2^2 + u_3^2 = c^2$.
- No physical effect is transmitted faster than light.
- The principle of relativity applies to all physical phenomena – only the relative motion of nonaccelerating observers can be detected by physical experiments.

The first assumption is equivalent to the assertion that if Newton's first law holds in one coordinate system, then it also holds in the other; in both, the worldlines of free particles – particles not acted on by any force – are straight lines in spacetime, given by linear equations. The second incorporates the assumption that the velocity of light should be independent of the observer. It must hold if light propagates by Maxwell's equations in an inertial coordinate system. The third assumption is needed for consistency, as we shall see.

We denote the top left entry in L by γ. This is the *time dilation factor* for the motion of O' relative to O. Along the worldline of O', which is given by $x' = y' = z' = 0$, we have

$$t = \gamma t' + \text{const.}$$

So γ relates the time measurements of events on the worldline of O' in the two coordinate systems. Similarly, if γ' is the top left entry in L^{-1}, then along the worldline of O $(x = y = z = 0)$ we have

$$t' = \gamma' t + \text{const.}$$

Hence γ' is the time dilation factor for the motion of O relative to O'. It follows from the fourth assumption, the relativity assumption, that the time dilation factor depends only on the relative motion of the two observers, and hence that $\gamma = \gamma'$.

It follows from the second assumption that the worldlines of photons through an event A form a cone in spacetime. This is called the *light cone* of the event.

If A is the event $t = x = y = z = 0$, then the light cone of A has the following equation

$$c^2 t^2 - x^2 - y^2 - z^2 = 0\,, \tag{3.7}$$

which is the condition that the time t elapsed from A at light speed should be related to the distance $D = \sqrt{x^2 + y^2 + z^2}$ from A by $D = ct$. The light cone consists of the event A itself, together with the *future light*

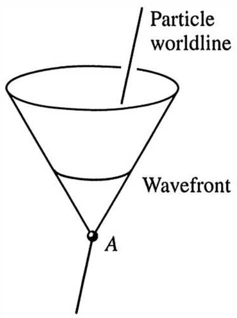


Fig. 3.7 The future light cone of A

cone, made up of the events that happen after A that can be reached from A by traveling at the speed of light; and the *past light cone*, made up of events that happen before A, from which A can be reached by traveling at the speed of light.

The future light cone of A is illustrated in the spacetime diagram (Fig. 3.7), where, as always, time runs up the page. The sections of the cone on which t is a positive constant are the spherical *wavefronts*, spreading out from the origin with speed c. By the third assumption, all particle worldlines through A must lie inside the light cone. We shall always draw spacetime diagrams so that the generators of the light cone are at 45° to the vertical.

All observers agree on the position in spacetime of the light cone of an event, so the cones determine an invariant structure on spacetime – a structure that was first made explicit by Minkowski. Thus the spacetime of special relativity is called *Minkowski space*.

3.3.2 Lorentz Transformation in Four Dimensions

Consider two events E_1 and E_2 with coordinates t_1, x_1, y_1, z_1 and t_2, x_2, y_2, z_2 in the first coordinate system set up by O; and with coordinates t'_1, x'_1, y'_1, z'_1 and t'_2, x'_2, y'_2, z'_2 in the second coordinate system set up by O'.

The two events lie on the worldline of a photon if and only if

$$\begin{aligned} &c^2(t_2 - t_1)^2 - (x_2 - x_1)^2 \\ &- (y_2 - y_1)^2 - (z_2 - z_1)^2 = 0\,, \end{aligned} \tag{3.8}$$

since this is the same as the condition $D = cT$, where

$$D = \sqrt{(x_2 - x_1)^2 - (y_2 - y_1)^2 - (z_2 - z_1)^2}$$

is the distance between them and $T = t_2 - t_1$ is time interval between them. Now *lying on the worldline of a photon* is a property that makes sense independently of any choice of coordinate system on spacetime. Hence if $D = cT$ according to one observer, then it must also be true according to the other. Therefore, (3.8) holds if and only if

$$\begin{aligned} &c^2(t'_2 - t'_1)^2 - (x'_2 - x'_1)^2 \\ &- (y'_2 - y'_1)^2 - (z'_2 - z'_1)^2 = 0\,. \end{aligned} \tag{3.9}$$

This statement can be written in a more compact form by putting

$$X = \begin{pmatrix} ct_2 \\ x_2 \\ y_2 \\ z_2 \end{pmatrix} - \begin{pmatrix} ct_1 \\ x_1 \\ y_1 \\ z_1 \end{pmatrix} \quad \text{and}$$

$$X' = \begin{pmatrix} ct'_2 \\ x'_2 \\ y'_2 \\ z'_2 \end{pmatrix} - \begin{pmatrix} ct'_1 \\ x'_1 \\ y'_1 \\ z'_1 \end{pmatrix}$$

and by defining g to be the 4×4 diagonal matrix

$$g = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix},$$

then (3.8) and (3.9) become, respectively,

$$X^{\mathrm{T}} g X = 0 \quad \text{and} \quad X'^{\mathrm{T}} g X' = 0\,,$$

where T denotes matrix transpose. Since the coordinates of the two events are related by (3.6), we have

$$X = LX'\,.$$

Therefore, the equivalence of (3.8) and (3.9) can be stated as follows. For $X \in \mathbb{R}^4$ (with the prime dropped),

$$X^{\mathrm{T}} g X = 0 \quad \text{if and only if} \quad X^{\mathrm{T}} L^{\mathrm{T}} g L X = 0\,.$$

It follows that $L^{\mathrm{T}} g L = \alpha g$ for some $\alpha \in \mathbb{R}$, which must be nonzero because the coordinate transformation must be nonsingular. Hence

$$L^{-1} = \alpha^{-1} g L^{\mathrm{T}} g\,,$$

since $g^{-1} = g$. Therefore, the top left entry in L^{-1} is γ/α, where γ is the top left entry in L. But we deduced from our relativity assumption that the top left entries in L and L^{-1} are equal, so $\alpha = 1$ and therefore $LgL^{\mathrm{T}} = g$.

3.3.3 Standard Lorentz Transformation

Suppose that O is moving along the x'-axis in the coordinates of O' and that O' is moving along the x-axis in the coordinates of O; also suppose further that they both take the origin of their coordinate systems to be the event at which they pass each other. Then $C = 0$ and the t, x and t', x' coordinates are related by (3.1), with $\gamma = \gamma(u)$. Hence

$$L = \begin{pmatrix} \gamma & \gamma\frac{u}{c} & p & q \\ \gamma\frac{u}{c} & \gamma & r & s \\ P & Q & a & b \\ R & S & c & d \end{pmatrix}.$$

From $L^{\mathrm{T}}gL = g$, we obtain

$$\gamma^2 - \gamma^2\frac{u^2}{c^2} - P^2 - R^2 = 1\,,$$

$$\gamma^2\frac{u^2}{c^2} - \gamma^2 - Q^2 - S^2 = -1\,.$$

But $\gamma^2 - \gamma^2 u^2/c^2 = 1$. Hence P, Q, R, and S are all zero. Similarly, from $LTgL = g$, we get that p, q, r, and s are also zero, and then that

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is an orthogonal matrix.

Although the direction of the x-axis in the coordinate system set up by O is fixed by the condition that O' should be traveling along the x-axis, O is still free to make rotations about the x-axis. By making an orthogonal transformation of the y and z coordinates by A^{-1}, it can be arranged without loss of generality that $A = 1$. We then have

$$\begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \gamma & \gamma\frac{u}{c} & 0 & 0 \\ \gamma\frac{u}{c} & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix}, \tag{3.10}$$

where $\gamma = \gamma(u) = 1/\sqrt{1 - u^2/c^2}$. This is the *standard Lorentz transformation* or *boost* with velocity u.

3.3.4 General Lorentz Transformation

In deriving the standard Lorentz transformation, we made assumptions about the relative orientations of the spatial axes of the two-coordinate systems. If we drop these, but still assume that O' is moving directly away from O, then we must combine (3.10) with an orthogonal transformation of the x, y, z coordinates and an orthogonal transformation of the x', y', z' coordinates. The result is

$$\begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} = L \begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} \tag{3.11}$$

with

$$L = \begin{pmatrix} 1 & 0 \\ 0 & H \end{pmatrix} L_u \begin{pmatrix} 1 & 0 \\ 0 & K^{\mathrm{T}} \end{pmatrix}, \tag{3.12}$$

where H and K are 3×3 proper orthogonal matrices, and L_u is the standard Lorentz transformation matrix with velocity u, for some $u < c$. A transformation of the form (3.11) is called a *proper orthochronous Lorentz transformation*. Such transformations are characterized by the following three properties:

1. The matrix L is *pseudo-orthogonal*. That is $L^{-1} = gL^{\mathrm{T}}g$.
2. The top left entry in L is positive.
3. $\det(L) = 1$.

A general Lorentz transformation is required to satisfy only (1). The second condition characterizes the transformation as *orthochronous*; that is, t is an increasing function of t'. If the third also holds, then L is *proper* and the handedness of the two sets of spatial axes is the same.

Finally, if we drop the condition that O' should be moving directly away from O, then we can combine (3.11) with a spatial translation and a change in the origin of the time coordinate. The result is

$$\begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} = L \begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} + C\,, \tag{3.13}$$

where L is a proper orthochronous Lorentz transformation matrix and C is a constant column vector. Equation (3.13) is an *inhomogeneous Lorentz transformation* or, alternatively, a *Poincaré transformation.*

In dealing with relativistic fields, it is conventional to put $x^0 = ct$, $x^1 = x$, $x^2 = y$, $x^3 = z$, and so on, and to

write (3.13) as

$$x^a = \sum_{b=0}^{3} L^a{}_b x'^b + C^a \quad (a = 0, 1, 2, 3) \tag{3.14}$$

or, with a summation convention for the repeated index b

$$x^a = L^a{}_b x'^b + C^a \, .$$

3.4 Vectors

In three-dimensional Euclidean space, a vector (or *three-vector*) X has three components that transform under rotation by

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = H \begin{pmatrix} X'_1 \\ X'_2 \\ X'_3 \end{pmatrix},$$

where X_1, X_2, X_3 are the components in the x, y, z coordinate system and X'_1, X'_2, X'_3 are the components in the x', y', z' coordinate system. A *four-vector* is similarly an object X that associates an element (X^0, X^1, X^2, X^3) of $\mathbb{R}^4$ with each inertial coordinate system.

The X^as $(a = 0, 1, 2, 3)$ are called the *components* of X. They are required to have the property that if two inertial coordinate systems are related by (3.13) then the components X^a in the first (unprimed) system are related to components X'^a in the second (primed) by

$$\begin{pmatrix} X^0 \\ X^1 \\ X^2 \\ X^3 \end{pmatrix} = L \begin{pmatrix} X'^0 \\ X'^1 \\ X'^2 \\ X'^3 \end{pmatrix} . \tag{3.15}$$

This rather awkward definition says no more than that a four-vector is an object with four components X^0, X^1, X^2, X^3 and that the components transform under an inhomogeneous Lorentz transformation of the coordinates by the associated linear transformation – the same transformation as the coordinates, but without the constant column vector. As with three-vectors, one can add four-vectors and take scalar multiples.

A key example is the *displacement vector* X from an event E_1 to an event E_2. If the events have respective coordinates t_1, x_1, y_1, z_1, and t_2, x_2, y_2, z_2 in some inertial coordinate system, then the displacement vector X from E_1 to E_2 has components

$$\begin{aligned} X^0 &= ct_2 - ct_1 \, , \quad & X^1 &= x_2 - x_1 \, , \\ X^2 &= y_2 - y_1 \, , \quad & X^3 &= z_2 - z_1 \, . \end{aligned} \tag{3.16}$$

This can be turned around into a more geometric definition of a four-vector, as an equivalence class of pairs of events, with two pairs equivalent whenever the quantities X^a in (3.16) are the same, in one and hence in every inertial coordinate system.

3.4.1 Temporal and Spatial Parts

If two inertial observers O and O' are at rest relative to each other, then their time axes in spacetime will be aligned and their inertial coordinate systems will be related by

$$\begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & H \end{pmatrix} \begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} + C \, , \tag{3.17}$$

where H is a 3×3 proper orthogonal matrix and C is a column vector – that is, by a rotation of the x, y, z axes, combined with a translation in space and time. In this special case, the components of a four-vector X in the two systems are related by

$$X^0 = X'^0 \, , \qquad \begin{pmatrix} X^1 \\ X^2 \\ X^3 \end{pmatrix} = H \begin{pmatrix} X'^1 \\ X'^2 \\ X'^3 \end{pmatrix} . \tag{3.18}$$

The time component is the same, while the three spatial components X^1, X^2, X^3 behave as the components of a three-vector $\boldsymbol{x}$. So we can decompose X into a *temporal part* X^0 and a *spatial part* $\boldsymbol{x}$ in a way that depends only on the velocity of the observer and not on the particular choice of origin and orientation of the spatial coordinate axes. The decomposition will be unchanged by the transformation between the coordinate systems of two observers at rest relative to each other. By contrast, under a general transformation between the inertial coordinate systems of two observers in relative motion, the direction of the time axis changes and the temporal and spatial parts are mixed up.

We shall write

$$X = (X^0, X^1, X^2, X^3)$$

as shorthand for X *has components* X^0, X^1, X^2, X^3 in a particular inertial coordinate system and

$$X = (\xi, \boldsymbol{x})$$

for X *has temporal part* ξ *and spatial part* $\boldsymbol{x}$ relative to a particular choice for direction in spacetime of the t-axis.

3.4.2 Inner Product

In Euclidean geometry, the distance between two points is determined by the dot product, which is an *inner product* on the space of three-vectors. If A, B are points in space and if $\boldsymbol{x}$ is the vector from A to B, then the distance from A to B is $\sqrt{\boldsymbol{x}.\boldsymbol{x}}$. The pseudo-orthogonality of Lorentz transformations leads to an analogous *indefinite* inner product on the space of four-vectors. That is, it has all the properties of an inner product, except that it is not positive definite.

The inner product $g(X, Y)$ of two four vectors X and Y is the real number

$$g(X, Y) = X^0Y^0 - X^1Y^1 - X^2Y^2 - X^3Y^3 ,$$

where X^a, Y^a, $a = 0, 1, 2, 3$, are the components of X and Y in an inertial coordinate system.

It follows from the first defining property of a Lorentz transformation that the definition does not depend on the choice of inertial coordinates.

The inner product is a symmetric bilinear form of the space of four-vectors. It can be written as

$$g(X, Y) = g_{ab}X^a Y^b , \tag{3.19}$$

where the g_{ab}s are the entries in the matrix g. That is

$$g_{00} = 1 , \quad g_{11} = g_{22} = g_{33} = -1 ,$$

and $g_{ab} = 0$ when $a \neq b$. In (3.19), there are two summations over the repeated indices $a, b = 0, 1, 2, 3$.

A further notational device is to put $X_a = g_{ab}X^b$, again with a summation over b. Then $X_0 = X^0$, $X_1 = -X^1, X_2 = -X^2, X_3 = -X^3$ and

$$\begin{aligned} g(X, Y) &= X_aY^a \\ &= X_0Y^0 + X_1Y^1 + X_2Y^2 + X_3Y^3 . \end{aligned} \tag{3.20}$$

The operation of forming the X_as from the components X^a of X is called *lowering the index*. The conventions for the positioning of indices are such that summations are always over one lower index and one upper index.

3.4.3 Classification of Four-Vectors

The fact that the inner product on four-vectors is not positive definite means that it is possible to distinguish between different types of four-vector according to the sign of the invariant $g(X, X)$. A four-vector X is said to be *timelike*, *spacelike*, or *null* as $g(X, X) > 0$, $g(X, X) < 0$, or $g(X, X) = 0$. Two four-vectors X and Y are *orthogonal* if $g(X, Y) = 0$.

A four-vector whose spatial part vanishes in some inertial coordinate system must be timelike; and a four-vector whose temporal part vanishes in some inertial coordinate system must be spacelike. Conversely, if X is timelike, then there exists an inertial coordinate system in which $X^1 = X^2 = X^3 = 0$. If X is spacelike, then there exists and inertial coordinate system in which $X^0 = 0$. The null vectors lie on the cone

$$(X^0)^2 - (X^1)^2 - (X^2)^2 - (X^3)^2 = 0 . \tag{3.21}$$

In the case of timelike and null vectors (but *not* spacelike vectors), the sign of the time component X^0 is invariant. A timelike or null vector X is said to be *future-pointing* if $X^0 > 0$ in some (and hence every) inertial coordinate system, and *past-pointing* if $X^0 < 0$. See Fig. 3.8, where the time axis is vertical and one spatial dimension is suppressed, and where *FPTL* denotes *future-pointing timelike*, and so on.

3.4.4 Causal Structure of Minkowski Space

In the case of displacement four-vectors, the classification has a direct interpretation in terms of the *causal structure* of Minkowski space. Suppose that E and F are events and that X is the displacement four-vector from E to F. In studying the *causal relationship* between E and F, we are interested in whether it is possible for

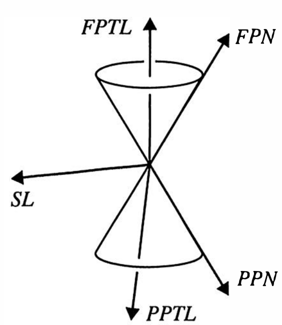


Fig. 3.8 The space of four-vectors

some physical process that happens at E to influence what happens at F, or the other way around.

The temporal part of X in an inertial coordinate system is the time from E to F multiplied by c; the spatial part is the three-vector from the point where E happens to the point where F happens. There are various possibilities.

Displacement X is Spacelike

In this case, it is impossible to get from E to F without traveling faster than light, so F lies outside the light cone of E, and vice versa. There exists an inertial coordinate system in which $X^0 = 0$; that is, in which E and F are simultaneous. If s denotes the distance from E to F in such coordinates, then

$$g(X,X) = -s^2 \,.$$

There also exist inertial coordinate systems in which E happens before F, and inertial coordinate systems in which E happens after F. It is for this reason that the prohibition on faster-than-light transmission is required for the consistency of the theory with commonsense ideas about causality.

Displacement X is Timelike

In this case there exists an inertial coordinate system in which $X^1 = X^2 = X^3 = 0$; that is in which E and F happen at the same place. If τ denotes the time from E to F in such coordinates, then

$$g(X,X) = c^2\tau^2 \,.$$

If X is future-pointing, then $\tau > 0$ and F happens after E in every inertial coordinate system. If X is past-pointing, then $\tau < 0$ and F happen before E in every inertial coordinate system.

Displacement X is Null

Then E and F lie on the worldline of a photon. If X is future-pointing (past-pointing), then F happens after (before) E in every inertial coordinate system.

3.4.5 Invariant Operators

In Euclidean space, the three partial derivatives with respect to Cartesian coordinates transform as the components of a vector operator ∇. By making ∇ act on a scalar field or a vector field, we can form the familiar invariant differential operators grad, div and curl.

There is an analogous *four-gradient*, which sends a function on spacetime to a four-vector field – a four-vector that varies from event to event – and *four-divergence*, which sends a four-vector field to a scalar function. To define them, put

$$\partial = (\partial_0, \partial_1, \partial_2, \partial_3) \,,$$

where $\partial_a = \partial/\partial x^a, a = 0, 1, 2, 3$. Now consider an inhomogeneous Lorentz transformation $x^a = L^a{}_b x'^b + C^a$. By the chain rule

$$\frac{\partial}{\partial x'^b} = \frac{\partial}{\partial x^a} L^a{}_b \,,$$

which, in matrix notation, is $\partial' = \partial L$.

This is very close to the transformation rule for a four-vector. By using the pseudo-orthogonality relation $L^{-1} = gL^{\mathrm{T}}g$, we can rewrite it in the form

$$\begin{pmatrix} \partial_0 \\ -\partial_1 \\ -\partial_2 \\ -\partial_3 \end{pmatrix} = L \begin{pmatrix} \partial'_0 \\ -\partial'_1 \\ -\partial'_2 \\ -\partial'_3 \end{pmatrix} , \tag{3.22}$$

where $\partial'_a = \partial/\partial x'^a$. Now the operator on the right transforms as a four-vector, which is called the *four-gradient*. It will be denoted by Grad.

Given a function f on spacetime, Grad$f$ is an intrinsically defined four-vector field. Similarly, given a four-vector field X, we can form an invariant scalar field Div$X$ by taking the inner product of the four-vector operator with X. The result is the *four-divergence* of a four-vector field X, defined by

$$\mathrm{Div}\, X = \frac{1}{c}\frac{\partial X^0}{\partial t} + \frac{\partial X^1}{\partial x} + \frac{\partial X^2}{\partial y} + \frac{\partial X^3}{\partial z} \,.$$

By combining these operations, we define the d'Alembertian, or wave operator, $\Box$, which acts on a function u by

$$\Box u = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} - \frac{\partial^2 u}{\partial z^2} = \mathrm{Div}(\mathrm{Grad}\, u) \,.$$

The d'Alembertian is an invariant operator. The result of applying it to a function is independent of the choice of inertial coordinates.

3.4.6 Frequency Four-Vector

The *wave equation* is the invariant condition $\Box u = 0$. A *real harmonic wave* is the real part of a complex solution of the form

$$\psi = A\exp(-\mathrm{i}\Omega) \,, \tag{3.23}$$

where $A = \alpha + \mathrm{i}\beta$ is constant and Ω is a real linear function of the inertial coordinates. Such complex solutions are characterized by the condition that

$$K = \frac{\mathrm{i}c\,\mathrm{Grad}\,\psi}{\psi}$$

should be a constant real null four-vector. It is called the *frequency four-vector*, and it has temporal and spatial parts

$$K = \omega(1, \boldsymbol{e}) ,$$

where ω is the frequency and $\boldsymbol{e}$ is the unit vector in the direction of propagation.

There is one subtlety here that is worth noting. If $u = \mathrm{Re}\psi$, then also $u = \mathrm{Re}\overline{\psi}$. Thus a real harmonic wave can be written in two distinct ways as the real part of a complex harmonic wave. In one case K is future-pointing, and in the other it is past-pointing. This is an important point in quantum field theory, where it is related to the distinction between particles and antiparticles. But in our classical context we shall avoid it by making the convention that ψ should always be chosen so that K is future-pointing.

3.5 Proper Time

The history or *worldline* of a particle in general motion is a curve in spacetime. If the particle is moving uniformly at constant speed, then its worldline Γ is a straight line which lies inside the light cone of any event on Γ.

We can label the events along such a straight worldline by using *proper time*, which is defined to be the time τ shown on a clock carried by the particle; τ is also the time coordinate in an inertial coordinate system set up by an observer moving with the particle, relative to whom the particle is at rest. It is a natural parameter, analogous to the distance along a line in Euclidean space. It is well-defined up to the addition of a constant, which is determined by the choice of the event $\tau = 0$.

If the particle is at rest in the inertial coordinate system $\tilde{t}, \tilde{x}, \tilde{y}, \tilde{z}$ then its worldline is given by $\tilde{t} = \tau$, with $\tilde{x}$, $\tilde{y}$, $\tilde{z}$ constant. If t, x, y, z is a second inertial coordinate system related to $\tilde{t}, \tilde{x}, \tilde{y}, \tilde{z}$ by the standard Lorentz transformation with velocity v, then

$$t = \gamma(v)\tau + \text{const.} \tag{3.24}$$

along Γ. Since the coordinate time is unchanged by rotation or translation, it follows that in a general inertial coordinate system t, x, y, z

$$\frac{\mathrm{d}t}{\mathrm{d}\tau} = \gamma(v) = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} , \tag{3.25}$$

where v is the speed of the particle. We see again the *time dilation* effect; when $v \neq 0$, a clock carried by the particle, which shows proper time, runs slow relative to the coordinate time t.

Along the particle worldline, the inertial coordinates are functions of τ. We put $V^a = \mathrm{d}x^a/\mathrm{d}\tau$. Then the V^as are the components of a four-vector V, called the *four-velocity* of the particle.

Suppose that the particle has velocity $\boldsymbol{v}$ relative to the inertial coordinate system t, x, y, z. Then $\mathrm{d}t/\mathrm{d}\tau = \gamma(v)$, where $v = |\boldsymbol{v}|$, and

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = \frac{\mathrm{d}t}{\mathrm{d}\tau}\frac{\mathrm{d}x}{\mathrm{d}t} = \gamma(v)v_1 , \tag{3.26}$$

and similarly for the other components. So V decomposes into temporal and spatial parts as

$$V = \gamma(v)\,(c, \boldsymbol{v}) . \tag{3.27}$$

It follows that $g(V, V) = c^2$. Thus, the four-velocity is a timelike four-vector. It is also future-pointing because $V^0 = c\gamma(v) > 0$.

3.5.1 Addition of Velocities

Suppose that an observer has velocity $\boldsymbol{u}$ and a particle has velocity $\boldsymbol{v}$ in some inertial coordinate system. And let w denote the speed of the particle relative to the observer.

If U and V denote the respective four-velocities of the observer and the particle, then in the given inertial coordinate system

$$U = \gamma(u)(c, \boldsymbol{u}) \quad \text{and} \quad V = \gamma(v)(c, \boldsymbol{v}) .$$

Hence

$$g(U, V) = \gamma(u)\gamma(v)(c^2 - \boldsymbol{u}.\boldsymbol{v}) .$$

On the other hand, in an inertial coordinate system in which the observer is at rest, $U = (c, 0)$ and $V =$

$\gamma(w)(c, \boldsymbol{w})$, where $\boldsymbol{w}.\boldsymbol{w} = w^2$. Hence

$$g(U, V) = \gamma(w) .$$

Therefore, $c^2\gamma(w) = \gamma(u)\gamma(v)(c^2 - \boldsymbol{u}.\boldsymbol{v})$. On solving for w, we find that

$$w = \frac{c\sqrt{c^2(\boldsymbol{u}-\boldsymbol{v}).(\boldsymbol{u}-\boldsymbol{v}) - u^2v^2 + (\boldsymbol{u}.\boldsymbol{v})^2}}{c^2 - \boldsymbol{u}.\boldsymbol{v}}$$

which reduces to the classical formula $w = |\boldsymbol{u}-\boldsymbol{v}|$ when $u, v \ll c$.

3.5.2 Lorentz Contraction

Suppose that a rod is of length L_0 in its rest frame and that in a second inertial frame, it is oriented in the direction of the unit vector $\boldsymbol{e}$ and is moving with velocity $\boldsymbol{v}$.

Let V denote the four-velocity of the rod. Let A be an event at one end of the rod, and let B and C be events at the other end such that B is simultaneous with A in the frame in which the rod is moving and C is simultaneous with A in the rest frame of the rod. The problem is to find the distance L between A and B, measured in the frame in which the rod is moving.

Let X be the displacement four-vector from A to B and let Y be the displacement four-vector from A to C. Then $Y = X + \tau V$ for some scalar τ and

$$L^2 = -g(X, X) , \quad L_0^2 = -g(Y, Y) .$$

Also $g(V, Y) = 0$ since A and C are simultaneous in the rest frame of the rod.

We find τ by calculating $g(X, V)$ in two different ways. In the frame in which the rod is moving, we have $X = (0, L\boldsymbol{e})$ and $V = \gamma(v)(c, \boldsymbol{v})$, so

$$g(X, V) = -L\gamma(v)\boldsymbol{e}.\boldsymbol{v} = -Lv\gamma(v)\cos\theta .$$

We also have $g(X, V) = g(Y - \tau V, V) = -\tau c^2$. Hence $\tau = Lv\gamma(v)\cos\theta/c^2$ and therefore

$$\begin{aligned} L^2 &= -g(Y - \tau V, Y - \tau V) \\ &= L_0^2 - \tau^2 c^2 \\ &= L_0^2 - c^{-2}L^2\gamma(v)^2v^2\cos^2\theta . \end{aligned}$$

After a little algebra, therefore, the length of the rod in the second frame is

$$L = \frac{L_0\sqrt{c^2 - v^2}}{\sqrt{c^2 - v^2\sin^2\theta}} ,$$

where θ is the angle between $\boldsymbol{e}$ and $\boldsymbol{v}$. This is the general formula for the Lorentz contraction.

3.6 Four-Acceleration

The worldline of an accelerating particle can be represented in inertial coordinates by equations of the form

$$x = x(t) , \quad y = y(t) , \quad z = z(t) ,$$

where t, x, y, z are inertial coordinates. Let E at time t and E' at time $t + \delta t$ be two nearby events on the worldline. In this general case, we define the *proper time* from E to E' to be the time $\delta\tau$ measured in a second inertial coordinate system in which the particle is instantaneously at rest at the event E.

In the t, x, y, z coordinate system, the displacement four-vector X from E to E' is $X = (c, \boldsymbol{v})\delta t$ where $\boldsymbol{v}$ is the velocity of the particle. From the discussion in Sect. 3.4.4

$$c^2\delta\tau^2 = g(X, X) = (c^2 - \boldsymbol{v}.\boldsymbol{v})\delta t^2 . \tag{3.28}$$

As in the case of a nonaccelerating particle, therefore, we define the proper time along the worldline up to an additive constant by $\mathrm{d}\tau/\mathrm{d}t = 1/\gamma(v)$ where v is the speed of the particle. The *Clock Hypothesis* asserts that proper time is the time measured by an ideal clock traveling with the particle, that is by a clock that is unaffected by the acceleration.

The four-velocity V is defined in the same way as for uniform motion, with the components $V^a = \mathrm{d}x^a/\mathrm{d}\tau$. Again $V = \gamma(v)(c, \boldsymbol{v})$, where $\boldsymbol{v}$ is the velocity measured in an inertial coordinate system and $v = |\boldsymbol{v}|$. But now V now depends on τ, and its derivative A with respect to τ is also a four-vector, called the. *four-acceleration.* In components, $A^a = \mathrm{d}V^a/\mathrm{d}\tau$.

In a general inertial coordinate system, A has spatial and temporal parts

$$\begin{aligned} A &= \gamma(v)\frac{\mathrm{d}}{\mathrm{d}t}(\gamma(v)(c, \boldsymbol{v})) \\ &= \frac{v\gamma(v)^4}{c^2}\frac{\mathrm{d}v}{\mathrm{d}t}(c, \boldsymbol{v}) + \gamma(v)^2\left(0, \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t}\right) . \end{aligned} \tag{3.29}$$

If the particle is instantaneously rest, then $V = (c, 0)$ and $A = (0, \boldsymbol{a})$ where $\boldsymbol{a} = \mathrm{d}\boldsymbol{v}/\mathrm{d}t$ is the ordinary acceleration. It follows that

$$g(A, V) = 0 , \quad g(V, V) = c^2 , \quad g(A, A) = -a^2 ,$$

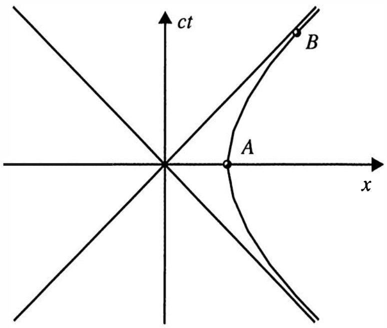


Fig. 3.9 A constant acceleration worldline

where a is the magnitude of the acceleration measured in the instantaneous rest frame; that is, the acceleration *felt* by an observer moving with the particle. A little further work establishes that $a = c\mathrm{d}\phi/\mathrm{d}\tau$, where ϕ is the rapidity determined by v.

3.6.1 Constant Acceleration

Suppose that $y = z = 0$ along the worldline in some fixed inertial coordinate system, and that a is constant. The components of V and A in the fixed coordinate system are $(c\dot{t}, \dot{x}, 0, 0)$ and $(c\ddot{t}, \ddot{x}, 0, 0)$ where the dot denotes $\mathrm{d}/\mathrm{d}\tau$, differentiation with respect to proper time. Hence

$$c^2\dot{t}^2 - \dot{x}^2 = c^2 \quad \text{and} \quad c^2\ddot{t}^2 - \ddot{x}^2 = -a^2 .$$

With a suitable choice of origin for the coordinates, and of the zero for τ, these integrate to give

$$t = \frac{c}{a}\sinh\left(\frac{a\tau}{c}\right) \quad \text{and} \quad x = \frac{c^2}{a}\cosh\left(\frac{a\tau}{c}\right).$$

So the worldline is a hyperbola with asymptotes $ct = \pm x$ (Fig. 3.9). Although at each event the acceleration measured in an inertial coordinate system in which the particle is instantaneously at rest is always the same, the velocity of the particle never exceeds c.

The coordinate time and the coordinate distance between the events at which $\tau = 0$ and at which τ takes some general value is

$$t = \frac{c}{a}\sinh\left(\frac{a\tau}{c}\right) \quad \text{and} \quad x = \frac{c^2}{a}\left(\cosh\left(\frac{a\tau}{c}\right) - 1\right).$$

By fixing τ and choosing a large value of a, we can make x as large as we please. Therefore, the prohibition on faster than light travel does not preclude travel over an arbitrarily large distance in a given interval of time. But the distance must be measured in the rest frame at $\tau = 0$, and time measured along the worldline. The catch is that if this is exploited for interstellar travel, starting and ending on earth, then the time that passes on earth before the completion of the journey is at least D/c, where D is the total distance traveled, as measured from earth. A traveler can complete a round trip journey of thousands of light years in a few years, as measured on the his own clock, but on his return thousands of years will have passed on earth.

There is an asymmetry between the traveler and the earth because the traveler is accelerating, while the earth is not, at least not significantly. It is sometimes said incorrectly that there is a paradox here (the *twin paradox*, since it is stated in terms of two twins, one in a spaceship, the other on earth). But the result is only paradoxical if one forgets that although uniform motion has only a relative meaning in special relativity, acceleration is absolute.

3.7 Visual Observation

The language of special relativity can sometimes mislead. For example, the statement *a measuring rod appears to an observer moving in the direction of the rod to have contracted* is true only if the phrase *appears ... to an observer* is interpreted in terms of a particular measuring procedure; the observer must set up inertial coordinates and then determine the distance between the worldlines of the two ends of the rod. It is tempting to make the erroneous assumption that other classically equivalent measurements will give the same result – for example, a measurement of the angle subtended by the two ends of the rod at known distance. In fact it does not because it involves *visual observation*; the motion of the observer also affects the measured angle between the trajectories of the photons arriving at the observer from the two ends of the rod. In fact, a moving rod does not even *appear* to be straight when observed visually.

To understand what an inertial observer actually sees at a particular event E on his worldline, we must consider the photon worldlines that pass through E. These are the generators of the light cone of E. An event A is on the past light cone of E if the displacement

four-vector K from A to E is null and future-pointing. In the observer's inertial frame, it has temporal and spatial parts

$$K = (\kappa, \boldsymbol{k}) \ ,$$

where $\kappa = \sqrt{\boldsymbol{k}.\boldsymbol{k}} > 0$. Light emitted at A arrives at the observer from the direction of the unit vector $-\boldsymbol{k}/\kappa$.

Suppose that B is a second event on the past light cone of E and that the displacement vector from B to E is the null four-vector $L = (\lambda, \boldsymbol{\ell})$. If the angle between the directions from which light from A and B arrives at the observer is θ, then

$$\cos\theta = \frac{\boldsymbol{k}.\boldsymbol{\ell}}{\kappa\lambda} = 1 - \frac{c^2 g(K, L)}{g(K, V)g(L, V)} \ , \qquad (3.30)$$

where V is the four-velocity of the observer. The second equality follows from the fact that $V = (c, 0)$ in the observer's frame. Therefore,

$$g(V, K) = c\kappa \ ,$$
$$g(V, L) = c\lambda \ ,$$
$$g(K, L) = \kappa\lambda - \boldsymbol{k}.\boldsymbol{\ell} \ .$$

3.7.1 Stellar Aberration

An application is the formula for stellar aberration. Suppose that an observer measures the angle subtended by two distant stars to be θ. Then a second observer moving relative to the first with speed v directly away from one of the stars measures the angle to be θ', where

$$\cos\theta' = \frac{c\cos\theta - v}{c - v\cos\theta} \ .$$

To show this, we take A and B to be events at which light from the stars is emitted, and we suppose that the second observer moves relative to the first directly away from the star at A. Then we have

$$g(K, L) = \kappa\lambda(1 - \cos\theta) \ ,$$

where $K = (\kappa, \boldsymbol{k})$, $L = (\lambda, \boldsymbol{\ell})$ in the first observer's frame, and the second observer has velocity $v\boldsymbol{k}/\kappa$ relative to the first.

In the first observer's frame, the four-velocity of the second observer is given by

$$V = \gamma(v)(c, v\boldsymbol{k}/\kappa) \ .$$

Hence,

$$g(K, V) = \gamma(v)\left(c\kappa - \frac{v\boldsymbol{k}.\boldsymbol{k}}{\kappa}\right) = \kappa\gamma(v)(c - v)$$

and

$$g(L, V) = \gamma(v)\left(c\lambda - \frac{v\boldsymbol{k}.\boldsymbol{\ell}}{\kappa}\right)$$
$$= \lambda\gamma(v)(c - v\cos\theta) \ .$$

From above, therefore

$$\cos\theta' = 1 - \frac{c^2 g(K, L)}{g(K, V)g(L, V)} = \frac{c\cos\theta - v}{c - v\cos\theta} \ , \qquad (3.31)$$

after a little algebra. If $\theta \neq 0$, so that the stars are separated in the sky, then $\cos\theta' \to -1$ as $v \to c$. So as the second observer looks in the direction of his motion relative to the first and accelerates toward the velocity of light, all stars appear to move across the sky to positions directly ahead. This includes the stars that were initially behind him, but not directly behind him.

3.7.2 Appearance of a Moving Sphere

A striking example of the distinction between visual observation and coordinate measurement is provided by a moving sphere. Whatever the speed of the sphere, the visually observed outline is always circular, despite the fact that, according to coordinate measurements, it is squashed along the direction of its motion by the Lorentz contraction. (This was first pointed out, surprisingly late in the development of relativity, by *Penrose* in 1959 [3.4].)

To see this, consider the light rays reaching an observer at the origin from the visually observed outline of a stationary sphere. If the three-vector from the center of the sphere to the observer is $\boldsymbol{x}$ and if the sphere subtends an angle 2α at the observer, then light reaching the observer from the outline of the sphere will travel in the direction of one of the unit vectors $\boldsymbol{e}$ that satisfies

$$\boldsymbol{e}.\boldsymbol{x} = |\boldsymbol{x}|\cos\alpha \ . \qquad (3.32)$$

Consider a photon emitted from the sphere at the event A that reaches the observer at the event E at which $t = x = y = z = 0$. Suppose that the displacement four-vector K from A to E has temporal and spatial parts $(\kappa, \boldsymbol{k})$ in the observer's frame. Then $\kappa = \sqrt{\boldsymbol{k}.\boldsymbol{k}}$ since K is future-pointing and null. If the photon appears to the

observer to come from the outline of the sphere, then $\boldsymbol{e} = \boldsymbol{k}/\kappa$ satisfies (3.32). So we can characterize the *outline events* E by the condition

$$\kappa|\boldsymbol{x}|\cos\alpha - \boldsymbol{k}.\boldsymbol{x} = 0\,.$$

That is, $g(K, X) = 0$, where X is the spacelike four-vector with temporal and spatial parts $(|\boldsymbol{x}|\cos\alpha, \boldsymbol{x})$ in the observer's frame.

In the frame of another inertial observer at E, $X = (\xi', \boldsymbol{x}')$ and $K = (\kappa', \boldsymbol{k}')$, with $\kappa' = |\boldsymbol{k}'|$. Since $g(K, X)$ is invariant, we shall have for the *outline events*

$$\kappa'\xi' - \boldsymbol{k}.\boldsymbol{x}' = 0\,,$$

and hence the photons reaching the second observer from the visually observed outline of the sphere travel in the directions $\boldsymbol{e} = \boldsymbol{k}'/\kappa'$ that satisfy

$$\boldsymbol{e}.\boldsymbol{x}' = |\boldsymbol{x}'|\cos\alpha'\,,$$

where $\cos\alpha' = \xi'/|\boldsymbol{x}'|$. Therefore, they appear to the second observer to come from a sphere with center in the direction of $-\boldsymbol{x}'$, with outline subtending an angle $2\alpha'$. To a moving observer, a sphere still has the visual appearance of a sphere, but of a different size. It still has a circular outline, and does not *look* squashed. If the second observer moves with speed v parallel to $\boldsymbol{x}$, then

$$\cot\alpha' = \gamma(v)\left(\cot\alpha - \frac{v}{c}\operatorname{cosec}\alpha\right)\,,$$

so $\alpha' \to \pi$ as $v \to c$. As the observer accelerates away from the sphere, the outline grows until it fills the whole sky, apart from a small hole directly ahead.

3.7.3 Möbius Transformations

Another way to derive this result is to identify the sphere of null lines through the origin (the *sky*) with the Riemann sphere. The Möbius group, or projective general linear group $\mathrm{PGL}(2, \mathbb{C})$, acts this sphere, by identifying it with the extendedcomplex plane by stereographic projection. We write $\boldsymbol{e} = (x, y, z)$, with $x^2 + y^2 + z^2 = 1$, and put

$$\zeta(\boldsymbol{e}) = \frac{x + \mathrm{i}y}{1 - z}$$

for $z \neq 1$, and $\zeta(\boldsymbol{e}) = \infty$ when $z = 1$. The Möbius group M then becomes the group of transformations of the sphere given by

$$\zeta \mapsto \frac{a + b\zeta}{c + d\zeta}\,, \tag{3.33}$$

where $a, b, c, d \in \mathbb{C}$, with $ad - bc \neq 0$. Two such transformations with parameters a, b, c, d and a', b', c', d' are the same whenever

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \mu \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix}, \quad \mu \neq 0 \in \mathbb{C}\,.$$

Lorentz transformations also act on the sphere by mapping null lines through the origin to null lines. These maps coincide with Möbius transformations and indeed we obtain in this way an isomorphism of the proper orthochronous Lorentz group with the group of Möbius transformations. A rotation through θ about the z-axis coincides with the Möbius transformation $\zeta \mapsto \mathrm{e}^{\mathrm{i}\theta}\zeta$. The permutation of the spatial coordinates $(x, y, z) \mapsto (z, x, y)$ coincides with

$$\zeta \mapsto \frac{\zeta + \mathrm{i}}{\zeta - \mathrm{i}}\,.$$

Finally, a Lorentz transformation with rapidity ϕ in the t, z-plane becomes, after stereographic projection from $(0, 0, 1)$, the dilation $\zeta \mapsto \mathrm{e}^{\phi}\zeta$. These transformations generate the whole proper orthochronous component of the Lorentz group. The correspondence between the actions on the sphere gives an homomorphism, which can easily be seen to be an isomorphism.

Because Möbius transformations map circles to circles, Lorentz transformations also map a circular outline on the sky to circular outlines.

3.8 Operational Definition of Mass

A central prediction of special relativity is the equivalence of mass and energy. The energy content of a body is determined by its dynamical mass, and the total amount of energy that can be released from a body is limited by the celebrated formula $E = mc^2$. By contrast, in classical physics there is in principle no limit to the amount of energy that can be stored in a body of given mass.

As with other relativistic predictions, the terms must be given operational definition. Mass and energy are quantities defined by the thought experiments used to measure them. They are not quantities with self-evident meaning taken over from classical theory.

Mass enters Newtonian mechanics in two ways, as *inertial mass*, in the second law $F = ma$, and as *gravitational mass*, in the inverse-square law $F = Gmm'/r^2$. It is the latter that one measures by weighing a body. But this is not a good starting point for our operational definition since there is no sensible way to include gravitational interactions in special relativity. Instead we begin with a direct *dynamical* measurement of mass in collisions.

The starting point is that the Newtonian conservation laws hold to a high degree of accuracy in collisions in which velocities of particles are much less than the velocity of light. Given a standard mass M, therefore, an observer can assign a mass m to any other particle by colliding it at low speed with the standard mass, measuring the resulting velocities, and applying the Newtonian law of conservation of momentum. Since this becomes exact as the velocities go to zero, the observer can in principle use a limiting procedure to measure m when the particle is at rest.

The rest mass of a particle is the mass measured by low-speed collisions in an inertial coordinate system in which the particle is at rest.

Rest mass is an intrinsic quantity associated with a particle.

3.8.1 Conservation of Four-Momentum

Each particle in a collision has a rest mass m (a scalar) and a four-velocity V (a four-vector). The four-vector $P = mV$ is called the *four-momentum* of the particle. It has temporal and spatial parts

$$P = (m\gamma(v)c, m\gamma(v)\boldsymbol{v}) \ ,$$

where $\boldsymbol{v}$ is the three-velocity. As $\boldsymbol{v} \to 0$, $\gamma(v) = 1 + O(v^2/c^2)$ and

$$P = (mc, m\boldsymbol{v}) + O\left(\frac{v^2}{c^2}\right) \ .$$

So if all the velocities are so small that terms in v^2/c^2 can be neglected, then the Newtonian laws of conservation of mass and momentum are equivalent to the conservation of the temporal and spatial parts of four-momentum.

We need to replace the Newtonian laws by statements that are equivalent when v^2/c^2 can be neglected, but which are otherwise compatible with Lorentz transformations. A very straightforward possibility is to adopt the hypothesis that four-momentum is always conserved.

Four-Momentum Hypothesis. *If the incoming particles in a collision have four-momenta $P_1, P_2, \ldots, P_k$ and the outgoing particles have four-momenta $P_{k+1}, P_{k+2}, \ldots, P_n$, then*

$$\sum_1^k P_i = \sum_{k+1}^n P_i \ . \tag{3.34}$$

The justification for this is, first, that it is equivalent to the Newtonian laws of conservation of mass and momentum for low-speed collisions and, second, since it is a relationship between four-vectors, it is compatible with Lorentz transformations; if it holds in one inertial frame, then it holds without approximation in every inertial frame.

Whatever the velocities of the particles, we can still take the temporal and spatial parts of (3.34) to obtain

$$\sum_1^k m_i\gamma(v_i) = \sum_{k+1}^n m_i\gamma(v_i) \ ,$$

$$\sum_1^k m_i\gamma(v_i)\boldsymbol{v}_i = \sum_{k+1}^n m_i\gamma(v_i)\boldsymbol{v}_i \ ,$$

where the m_i's are the rest masses of the particles. These take the same form as the Newtonian laws of mass and momentum conservation when we identify $m\gamma(v)$ with *inertial mass* and $m\gamma(v)\boldsymbol{v}$ with *three-momentum.*

Four-momentum conservation is equivalent to conservation of inertial mass and of three-momentum (in every inertial coordinate system). The new feature of the relativistic theory is that the inertial mass of a particle increases with its velocity, albeit only very slightly for velocities much less than that of light.

Rest mass is a scalar – by its operational definition, it is an intrinsic quantity. But inertial mass is different in different inertial coordinate systems. Rest mass and inertial mass are equal for a particle at rest.

3.8.2 Photons

Alternative operational definitions of mass and energy start from the quantum mechanical principle that

a photon with (angular) frequency ω carries momentum $\hbar\omega/c$ in the direction of its motion. So if a particle at rest acquires small velocity v through the absorption of a low-energy photon at some event, then we can use

$$mv = \frac{\hbar\omega}{c}$$

to determine its rest mass m after the event. If we look at this in a frame in which the particle is at rest after the event, then we have a particle with low speed v being brought to rest by absorbing a photon with small momentum $\hbar\gamma(v)(1-v/c)/c$. So, because $\gamma \approx 1$ for low speeds, the rest mass of the particle before the collision is given by

$$m'v = \frac{\hbar\omega(c-v)}{c^2} .$$

The absorption of the photon has increased the rest mass of the particle. By putting the two equations together, we have

$$m'(c,0) + \frac{\hbar\omega}{c}(1,1) = m(c,v) .$$

Again using the approximation $\gamma(v) \approx 1$, this is the four-vector equation

$$m'V' + \frac{\hbar}{c}K = mV,$$

where K is the frequency four-vector of the photon. Thus we again have the four-momentum conservation law, provided that we assign four-momentum $\hbar K/c$ to the photon. It must hold for small ω. But by additivity, it must also hold for arbitrary ω, if we assume that n photons with frequency ω traveling in the same direction have the same total energy and momentum as a single photon of frequency $n\omega$.

3.8.3 Equivalence of Mass and Energy

In a general collision, it is not rest mass that is conserved, but the temporal part

$$P^0 = m\gamma(v)$$

of the four-momentum. Now

$$\gamma(v) = 1 + \frac{v^2}{2c^2} + O\left(\frac{v^4}{c^4}\right) .$$

So if we neglect terms of order v^4/c^4, but keep terms of order v^2/c^2, then

$$P^0 = \frac{1}{c}\left(mc^2 + \tfrac{1}{2}mv^2\right) ,$$

where m is the rest mass. Thus cP^0 is the sum of the Newtonian kinetic energy and a much larger term mc^2, which also has the dimensions of energy.

For a particle of rest mass m, the quantity $E = mc^2$ is the rest energy of the particle.

Any collision that involves a gain or loss of kinetic energy, such as an explosion or an inelastic collision, must involve a corresponding loss or gain in the total rest energies of the particles; kinetic energy can be traded for rest mass, and vice versa.

3.9 Maxwell's Equations

The requirement that Maxwell's equations should be consistent with the principle of relativity implies that the velocity of photons must be independent of the motion of their source and of the observer. That, in conjunction with other plausible assumptions leads to the conclusion that inertial coordinate systems must be related by Lorentz transformations. It is not immediately obvious, however, that this chain of reasoning is reversible, and that Maxwell's equations are in fact invariant. To show that, we must find the transformation rule for the components of the electric and magnetic fields. We must address the question: if an observer moves with velocity $\boldsymbol{v}$ through a given electromagnetic field, what electric and magnetic fields will he observe and do the observed fields satisfy Maxwell's equations?

We answer the first part of the question by considering observations of the motion of a charged particle relative to an inertial frame. The path of a particle moving slowly through an electric field $\boldsymbol{E}$ and magnetic field $\boldsymbol{B}$ is determined by the Lorentz force law. If the particle has velocity $\boldsymbol{v}$, momentum $\boldsymbol{p}$, and charge e, then

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = e(\boldsymbol{E} + \boldsymbol{v} \wedge \boldsymbol{B}) . \tag{3.35}$$

By measuring the trajectory for different velocities, an observer can in principle determine $\boldsymbol{E}$ and $\boldsymbol{B}$ at each point. Knowing how to apply an arbitrary Lorentz transformation to $\boldsymbol{E}$ and $\boldsymbol{B}$ is equivalent to knowing how to extend the equation of motion (3.35) to any $\boldsymbol{v}$ with $|\boldsymbol{v}| < c$. For if we know how to do the former, then we can transform to a frame in which the particle is moving slowly, find its trajectory, and then transform back to the original coordinates.

The transformation law for $\boldsymbol{E}$ and $\boldsymbol{B}$ must correctly encode the behavior of particles moving at high speed through electric and magnetic fields. This has two features (amply verified in particle accelerators). First, the rest mass m and charge e of a particle are unchanged by interaction with the fields; and, second, if we take $\boldsymbol{p}$ to be the spatial part of the four-momentum, then (3.35) holds for any $\boldsymbol{v}$ with $|\boldsymbol{v}| < c$. That is, the motion of a charged particle at any velocity is governed by

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = e(\boldsymbol{E} + \boldsymbol{v} \wedge \boldsymbol{B}) , \quad \text{where} \quad \boldsymbol{p} = m\gamma(v)\boldsymbol{v} . \tag{3.36}$$

These can be recast in terms of the temporal and spatial parts of the four-acceleration $A = (\alpha, \boldsymbol{a})$ as

$$mc\alpha = e\gamma(v)\boldsymbol{E}.\boldsymbol{v} , \quad m\boldsymbol{a} = e\gamma(v)(\boldsymbol{E} + \boldsymbol{v} \wedge \boldsymbol{B}) .$$

Since m is constant and $\boldsymbol{p} = m\gamma(v)\boldsymbol{v}$, the second equation follows from (3.36), together with

$$\boldsymbol{a} = \gamma(v)\frac{\mathrm{d}}{\mathrm{d}t}(\gamma(v)\boldsymbol{v}) .$$

The first follows from the orthogonality of the four-acceleration $A = (\alpha, \boldsymbol{a})$ and the four-velocity $V = \gamma(v)(c, \boldsymbol{v})$, which implies that $c\alpha = \boldsymbol{a}.\boldsymbol{v}$.

The transformation of the fields now follows by introducing the electromagnetic field tensor, which is represented by the 4×4 matrix

$$F = \begin{pmatrix} 0 & E_1 & E_2 & E_3 \\ -E_1 & 0 & -cB_3 & cB_2 \\ -E_2 & cB_3 & 0 & -cB_1 \\ -E_3 & -cB_2 & cB_1 & 0 \end{pmatrix} . \tag{3.37}$$

The entries in F are denoted by F_{ab}, $a, b = 0, 1, 2, 3$. They are determined by measuring the acceleration of a charged particle.

The term *tensor* refers to the behavior of F under Lorentz transformations: if the inertial coordinate systems of two inertial observers are related by (3.13), then the electromagnetic fields measured by the two observers at an event are related by $F' = L^{\mathrm{T}}FL$. This follows by writing the equation of motion in the matrix notation as

$$cmA = egFV .$$

where, as usual, g is the diagonal matrix with diagonal entries $1, -1, -1, -1$. But four-velocity and four-acceleration transform as four-vectors, so $V = LV'$ and $A = LA'$; and by the pseudo-orthogonality property, we have $L^{-1} = gL^{\mathrm{T}}g$. Hence

$$cmLA' = egFLV'$$

and therefore

$$cmA' = egL^{\mathrm{T}}FLV'$$

since g^2 is the identity. But in the second coordinate system

$$cmA' = eF'gV' .$$

Since both equations for A' hold whatever the four-velocity, we conclude that $F' = L^{\mathrm{T}}FL$.

Note that the transformation preserves the skew-symmetry of F. It should also be remarked that it is assumed implicitly that the charge of a particle is the same in all inertial coordinate systems. One piece of physical evidence for this is the overall neutrality of matter. When at rest, electrons and protons have equal and opposite charges. In an atom the electrons are moving much faster than the protons in the nucleus. If the charge of a particle depended on its velocity, then there could not be an exact balance between the electric charges of the electrons and the protons.

More generally, a tensor of type (p, q) is an object that associates components $T^{a\ldots c}{}_{e\ldots f}$ (p upper indices, q lower indices) with each inertial coordinate system, subject to the transformation rule

$$T^{a\ldots c}{}_{e\ldots f}L^{e}{}_{p}\ldots L^{f}{}_{r} = L^{a}{}_{s}\ldots L^{c}{}_{u}T'^{s\ldots u}{}_{p\ldots r}$$

under (3.13). An electromagnetic field is a tensor of type $(0, 2)$; a four-vector is a tensor of type $(1, 0)$. The coefficients g_{ab} in the definition of the inner product are also the components of a tensor of type $(0, 2)$, but one that is exceptional in having the same components in every inertial coordinate system. It is called the *metric tensor*.

3.9.1 Transformations of E and B

In the case of the standard Lorentz transformation,

$$L = \begin{pmatrix} \gamma & \gamma\frac{u}{c} & 0 & 0 \\ \gamma\frac{u}{c} & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

where $\gamma = \gamma(u)$, the tensor property $F' = L^{\mathrm{T}}FL$ gives

$$E'_1 = E_1\,, \quad E'_2 = \gamma(E_2 - uB_3)\,, \quad E'_3 = \gamma(E_3 + uB_2) \tag{3.38}$$

and

$$B'_1 = B_1\,, \quad B'_2 = \gamma(B_2 + uE_3/c^2)\,, \quad B'_3 = \gamma(B_3 - uE_2/c^2) \tag{3.39}$$

so the transformation mixes electric and magnetic fields. If $u \ll c$, then

$$\boldsymbol{E}' = \boldsymbol{E} + \boldsymbol{u} \wedge \boldsymbol{B}\,,$$

where $\boldsymbol{u} = u\boldsymbol{i}$. Thus an observer moving slowly with velocity $\boldsymbol{u}$ through a pure magnetic field $\boldsymbol{B}$ with $\boldsymbol{E} = 0$ sees an electric field $\boldsymbol{u} \wedge \boldsymbol{B}$.

It follows from (3.38) and (3.39) that $\boldsymbol{E}' \cdot \boldsymbol{B}' = \boldsymbol{E} \cdot \boldsymbol{B}$. Hence $\boldsymbol{E}.\boldsymbol{B}$ is invariant under standard Lorentz transformations. Clearly it is also invariant under rotations. Hence it is an invariant of the electromagnetic field – it is the same in every inertial coordinate system. Another invariant is $\boldsymbol{E}\cdot\boldsymbol{E} - c^2\boldsymbol{B}\cdot\boldsymbol{B}$. It follows, for example, that if $\boldsymbol{E}$ and $\boldsymbol{B}$ are not orthogonal in some inertial coordinate system, then $\boldsymbol{E} \neq 0$ and $\boldsymbol{B} \neq 0$ in every inertial coordinate system.

3.9.2 Invariance of Maxwell's Equations

In terms of the electromagnetic field, Maxwell's equations are

$$\partial_a F_{bc} + \partial_b F_{ca} + \partial_c F_{ab} = 0\,,$$
$$\partial^a F_{ab} = \frac{1}{\epsilon_0 c} J_b\,,$$

where $\partial_a = \partial/\partial x^a$, the ∂^as are the components of the four-gradient

$$\mathrm{Grad} = (\partial^0, \partial^1, \partial^2, \partial^3) = (\partial_0, -\partial_1, -\partial_2, -\partial_3)\,,$$

and the J_as are defined in terms of the charge and current density by

$$J = (J_0, J_1, J_2, J_3) = (c\rho, -j_1, -j_2, -j_3)\,.$$

If the J_as transform by $J'_b = J_a L^a{}_b$, with summation, or $J' = JL$ in matrix notation, then it follows from the transformation rules for the components of Grad and F that Maxwell's equations are invariant. In fact the transformation law for charge and current density follows from the invariance of charge and the transformation of volumes.

In the language of tensor calculus, F is a skew-symmetric second-rank covariant tensor, or 2-form, and J is a covector field, or 1-form.

References

3.1 N.M.J. Woodhouse: *Special Relativity* (Springer, London 2003)

3.2 E.A. Milne: *Relativity, Gravitation and World-Structure* (Oxford Univ. Press, Oxford 1935)

3.3 H. Bondi: *Assumption and Myth in Physical Theory* (Cambridge Univ. Press, Cambridge 1967)

3.4 R. Penrose: The apparent shape of a relativistically moving sphere, Proc. Camb. Phil. Soc. **55**, 137–139 (1959)

4. Acceleration and Gravity: Einstein's Principle of Equivalence

Lewis Ryder

It is shown that Einstein's equivalence principle, formulated in a manner consistent with Special Relativity, is inconsistent with the field equations of General Relativity.

4.1 Prologue

The success of special relativity demonstrated that there is no such thing as absolute velocity; all velocities are relative. And coupled with the observation that the speed of light is absolute, Einstein showed that this implies that time, like space, is relative, and indeed that they are merely the three- and one-dimensional components of what we now call spacetime, the four-dimensional continuum introduced by Minkowski. Einstein then asked himself: Can there be any sense in which *acceleration* is relative? Can the laws of physics retain their form, or something like their form, if we pass from one (say, inertial) frame of reference to another one accelerating with respect to it? He noted that in an accelerating frame time is indeed affected, so that accelerating clocks go slower, but he also noted that, by virtue of Galileo's observation that falling bodies in a gravitational field all fall at the same rate, the same phenomenon (of clocks going slower) should be expected in a gravitational field. The clocks in a gravitational field will go slower than clocks in empty space, which implies a *gravitational redshift*. The equivalence of a gravitational field to an accelerating frame of reference is what is known as the equivalence principle. The above line of reasoning, however, also marked the beginning of *general* relativity: by generalizing the Lorentz transformations (which are linear) to nonlinear ones (like a transformation to an accelerating frame), we automatically bring in gravity, so a theory of general relativity is a theory of gravity. This is the vision, very broadly stated, though it turns out that there are many observations and qualifications to be made. Some of these are spelled out below.

4.2 The Role of the Equivalence Principle in General Relativity

Einstein was a master of the thought experiment and what he described as the happiest thought of his life is a most beautiful and arresting insight. In a letter to R.W. Lawson, 22 January 1920 (quoted in [4.1, p. 178]), he writes:

> *When, in 1907, I was working on a comprehensive paper on the special theory of relativity for the* Jahrbuch der Radioaktivität und Elektronik, *I had also to attempt to modify the Newtonian theory of gravitation in such a way that its laws would fit in the [special relativity] theory. Attempts in this direction did show that this could be done, but did not satisfy me because they were based on physically unfounded hypotheses... Then there occurred to me the* glücklichste Gedanke meines Lebens, *the*

> *happiest thought of my life, in the following form. The gravitational field has only a relative existence in a way similar to the electric field generated by magnetoelectric induction.* Because for an observer falling freely from the roof of a house there exists – *at least in his immediate surroundings* – no gravitational field *[Einstein's emphasis]. Indeed, if the observer drops some bodies then these remain in a state of rest or of uniform motion, independent of their particular chemical or physical nature (in this consideration the air resistance is, of course, ignored). The observer therefore has the right to interpret his state as at rest. Because of this idea, the uncommonly peculiar experimental law that in a gravitational field all bodies fall with the same acceleration attained at once a deep physical meaning. Namely, if there were to exist just one single object that falls in a gravitational field in a way different from all others, then with its help the observer could realise that he is in a gravitational field and is falling in it. If such an object does not exist, however – as experience has shown with great accuracy – then the observer lacks any independent means of perceiving himself as falling in a gravitational field. Rather he has the right to consider his state as one of rest and his environment as field-free relative to gravitation.*
>
> *The experimentally known matter independence of the acceleration of fall is therefore a powerful argument for the fact that the relativity postulate has to be extended to coordinate systems which, relative to each other, are in nonuniform motion.*

Imagine, then, an observer standing on the floor of a closed box with no windows, who, releasing objects of different weights and made of different materials, observes that they fall at the same rate. He will conclude either that he is in a gravitational field (for instance that of the earth) or that the box is in no gravitational field, but somewhere in space, far from heavy planets or stars, and is being accelerated. In this last case, when objects are released, they fall to the ground with the same acceleration since it is the floor that accelerates up to meet them. A gravitational field is therefore indistinguishable from an accelerating frame. Einstein concluded from this observation, making at the same time a bold generalization, that *no experiment in mechanics* could distinguish a gravitational field from an accelerating frame. And of course, precisely because of this equivalence, a gravitational field may be *annulled* by acceleration – exactly Einstein's happiest thought, recalled above. In the case of the lift, free fall results if the suspending cable is severed. But now note that if the lift shaft were extremely deep, extending a significant distance into the earth's interior, then objects at different places in the lift would *move toward each other*, since they each are traveling along a radius vector toward the center of the earth. In this case the cancelation of the gravitational field by an accelerating frame is not complete. Differently stated, a *uniform* gravitational field would indeed be indistinguishable from an accelerating frame, but a realistic, physical one, such as that of the earth or the sun, would not. The equivalence principle is therefore a *local* principle, which may be stated by saying that a uniform gravitational field is indistinguishable from an accelerating frame. This formulation is all right as far as it goes, but of course in nature there is no such thing as a uniform gravitational field. Most gravitational fields are, in some approximation, radial. To retain a point of contact with the real world, then, let us simply consider actual gravitational fields (for example that of the earth), but limit our attention to a region small enough that the field is approximately uniform. Finally, to this restriction to locality may be added another consideration, which is to *generalize* the equivalence we are thinking about from mechanics to *all the laws of physics*; and the equivalence principle may then be stated:

> *In a freely falling (nonrotating) laboratory occupying a small region of spacetime, the local reference frames are inertial and the laws of physics are consistent with special relativity.*

Some writers distinguish two versions of the equivalence principle: the weak equivalence principle, which refers only to free fall in a gravitational field and is stated in [4.2, p. 1050] as *The worldline of a freely falling test body is independent of its composition or structure*; and the strong equivalence principle, according to which no experiment in any area of physics should be able, locally, to distinguish a gravitational field from an accelerating frame. This distinction has its origin in the generalization considered above, but it is not clear that it is a helpful, or even a valid distinction. Bodies in free fall are, after all, made of atoms, assembled as some sort of condensed matter, in which electrodynamic and nuclear forces, together with their relevant binding energies, are inevitably involved. And indeed, on top of this, quantum mechanical, and even quantum field theoretic, considerations will come into play, so it would seem that it is difficult to escape from a situation in which many, or most, of the laws of

physics are being investigated, when we are investigating how bodies behave in a gravitational field. For this reason I prefer to view the equivalence principle simply as the statement above, and not distinguish a weak equivalence principle from a strong one.

It is helpful to make the ideas above more precise. In Einstein's 1907 review of relativity, referred to above, Einstein was considering the nature of space and time in a uniformly accelerated system, and noted that in this system the time parameter σ (Einstein's notation) is related to the time parameter τ in an inertial frame by [4.3]

$$\sigma = \tau\left(1 + \frac{ax}{c^2}\right), \tag{4.1}$$

where a is the acceleration (denoted by γ by Einstein), along the x-axis. Einstein noted that this relation is only approximate, holding only if x is below a certain limit. He then noted that if the accelerated system were placed, instead, in a gravitational field with potential $\phi = gx$, with g the acceleration due to gravity (the same for all bodies, à la Galileo), we should therefore expect the same equation to hold,

$$\sigma = \tau\left(1 + \frac{\phi}{c^2}\right). \tag{4.2}$$

A clock in a gravitational field would then show a reading $\left(1 + \frac{\phi}{c^2}\right)$ times what it would show in no field.

By 1911, Einstein had developed this idea: defining a system K at rest in a homogeneous gravitational field, and an accelerated system K', he writes [4.4]:

> *But we arrive at a very satisfactory interpretation of the empirical law if we assume that the systems K and K′ are, physically, perfectly equivalent, i. e. if we assume that the system K could likewise be conceived as occurring in a space free of a gravitational field; but in that case, we must consider K as uniformly accelerated. Given this conception, one can no more speak of the absolute acceleration of the reference system than one can speak of a system's absolute velocity in the ordinary theory of relativity. With this conception, the equal falling of all bodies in a gravitational field is self-evident.*

By 1916, this equivalence principle, between a homogeneous gravitational field and an accelerating frame, had become a corner stone of the fully fledged General Theory of Relativity [4.5]. This theory was cast, however, in an almost unimaginably different language, based on the idea that spacetime is a pseudo-Riemannian manifold, whose curvature manifests itself as gravitation. In this same year, Schwarzschild published his (vacuum) solution to the Einstein field equations, which showed g_{00}, the time–time component of the metric tensor, to have a form consistent with (4.2) above. With the square of the invariant spacetime separation between events, in spherical polar coordinates, written as

$$\begin{aligned} ds^2 &= g_{\mu\nu}\,dx^\mu\,dx^\nu \\ &= g_{00}c^2\,dt^2 + g_{11}\,dr^2 + g_{22}r^2\,d\theta^2 \\ &\quad + g_{33}r^2\sin^2\theta + g_{01}c\,dt\,dr + \dots . \end{aligned} \tag{4.3}$$

Schwarzschild found

$$g_{00} = -\left(1 + \frac{2\phi}{c^2}\right). \tag{4.4}$$

This solution is, however, as noted, *exact*, and moreover makes no appeal to the equivalence principle. From a logical point of view, therefore, the equivalence principle is dispensible; the general theory may be obtained simply by enlarging the hypothesis of special relativity, that spacetime is of the Minkowski form, to the hypothesis that it is (pseudo-)Riemannian. The Einstein field equations, which are differential equations for the metric tensor $g_{\mu\nu}$, then enable this tensor to be found. The equivalence principle has nothing to add to this.

Let us take stock of the situation. Einstein's happy thought suggested to him that nature does not distinguish between a gravitational field and an accelerating reference frame. This implied in turn that time goes slower in a gravitational field. In fact, there are immediately three types of experimental tests to verify the equivalence principle, and these will be considered in the next section. On the other hand, the equivalence principle is *local*, for, as we have noted, over longer distances, objects in free fall in a realistic gravitational field move toward each other, and this does *not* happen in an accelerating frame. This effect is called a *tidal effect*, and it turns out that it is accounted for in general relativity as being a consequence of the *curvature* of spacetime; this is something which goes beyond the equivalence principle. The equivalence principle might well have been (in fact, was!) a source of direct inspiration to Einstein *en route* to discovering a new theory of gravity, but this new theory – general relativity – was mathematically much more sophisticated than the equivalence principle; and, as far as we know, is an exact and complete theory of gravity, at least at the classical level.

From a fundamental point of view, therefore, and with the hindsight of general relativity, the equivalence principle may be regarded as irrelevant. One of the most ardent proponents of this view is *Synge*, who states in the introduction to his book [4.6]:

> ... *I have never been able to understand this [Equivalence] Principle. Does it mean that the signature of the spacetime metric is* +2 *(or* −2 *if you prefer the other convention)? If so, it is important, but hardly a Principle. Does it mean that the effects of a gravitational field are indistinguishable from the effects of an observer's acceleration? If so, it is false. In Einstein's theory, either there is a gravitational field or there is none, according as the Riemann tensor does or does not vanish. This is an absolute property; it has nothing to do with any observer's worldline.*

4.3 Experimental Tests

Consider once more a test body in free fall in the earth's gravitational field. Newton's law states that it is subject to a force

$$F=\frac{m_{\mathrm{g}}MG}{R^2}=m_{\mathrm{g}}g\,,$$

where m_{g} is the *gravitational mass* of the test body, M and R are the mass and radius of the earth, and g the acceleration due to gravity. Newton's second law of motion, on the other hand, says that a body subjected to a force F will move with an acceleration a given by

$$F=m_{\mathrm{i}}a\,,$$

where m_{i} is the *inertial mass* of a body (considered a constant). It is clear that the concepts of gravitational and inertial mass are independent. Equating these expressions gives

$$a=\frac{m_{\mathrm{g}}}{m_{\mathrm{i}}}g\,.$$

The observation that all bodies fall at the same rate then implies that $m_{\mathrm{g}}/m_{\mathrm{i}}$ is the same for all bodies. In fact, the definition of G is chosen so that $m_{\mathrm{g}}=m_{\mathrm{i}}$, and the first test of the equivalence principle is to verify this equality. Accurate experiments were performed by Eötvös in 1889 and 1908. He made a torsion balance, from which two masses, of (in his case) gold and aluminum were suspended. If $m_{\mathrm{g}}/m_{\mathrm{i}}$ is not the same for the two metals the sun will exert a torque on the balance, and 12 h later, with the sun in the opposite direction, the torque will likewise act oppositely, thus causing an oscillation of the balance with a period of 24 h. This was not observed, and Eötvös concluded that the ratio of gravitational to inertial masses for gold and aluminum did not differ from unity by more than five parts in 10^9. A more recent investigation, using beryllium and copper, gives [4.7]

$$\eta=(-0.2\pm2.8)\times10^{-12}\,,$$

where

$$\eta=\frac{\alpha_1-\alpha_2}{(\alpha_1+\alpha_2)/2}\,,$$

and

$$\alpha=\frac{m_{\mathrm{g}}}{m_{\mathrm{i}}}\,,$$

and the subscripts 1 and 2 refer to beryllium and copper, respectively. For more information on experiments testing the equality of gravitational and inertial masses, see [4.2, 8–10].

The other tests for the equivalence principle are based on (4.2) or (4.4) – that time goes slower (and therefore clocks go slower) in a gravitational field. The traditional test for this is the gravitational frequency shift, and the most convincing demonstration of this is the Pound–Rebka observation of a blue-shift of radiation travelling vertically downward in the earth's gravitational field. The effect is tiny but the observations are extremely accurate: the prediction is a fractional shift of 2.46×10^{-15}, against an observed value of $(2.57\pm0.26)\times10^{-15}$.

In more recent years, this type of test, based on clocks going slower in a gravitational field, has been tested at the atomic level. A recent paper, *A precision measurement of the gravitational redshift by the interference of matter waves*, claims, on the basis of laboratory experiments involving quantum interference of atoms, to have increased the accuracy from 7×10^{-5}, for the tower-based experiment described above, to 7×10^{-9} [4.11]. This finding, or rather its interpretation, has been challenged, however; [4.12] and the original authors have replied to this challenge [4.13]. For a good review of these matters see [4.14].

A remarkable verification of the reality of time dilation in a gravitational field is, however, the operation of the global positioning system (GPS). This is an array of 24 satellites, each in a 12 h orbit round the earth. Each satellite carries an atomic clock, and the purpose is to locate, to an accuracy of about 10 m, any point on the earth's surface. This is done by sending radio signals between the satellites and the receiver on the earth, with the times of transmission and reception recorded; it is then trivial to calculate the distances involved – and of course only three satellites are in principle needed to pinpoint the position of the receiver on the earth. The interesting and relevant point is that the relative nature of time must be taken into account – arising both from special relativity (the satellites are moving) and from the equivalence principle (clocks on the earth go slower than those in the satellites). If these factors are not taken into account, the system breaks down in a matter of hours.

It is useful to illustrate this with some figures. The 24 satellites describe an orbit of radius 27 000 km and so are 7000 km apart, and by virtue of orbiting every 12 h travel at about $4\,\mathrm{km\,s^{-1}}$ – fast enough for special relativity to be relevant. Clocks register proper time, and for a moving clock this is γt,

$$\gamma = \left(1 - \frac{v^2}{c^2}\right)^{-1/2} \approx 1 + \frac{v^2}{2c^2}\,.$$

For a clock in an orbit of radius R, a Newtonian calculation gives $v^2 = MG/R$ so the special relativistic correction is of order

$$\frac{MG}{2Rc^2} \approx 0.8 \times 10^{-10}\,.$$

The gravitational time correction is, from (4.2),

$$\frac{\phi}{c^2} = \frac{GM}{Rc^2} \approx 1.6 \times 10^{-10}\,,$$

twice as large as the velocity correction factor. Hence in every second the clocks need to be adjusted by about 10^{-10} s. This might not seem much, but is certainly relevant for the accuracy needed – and is well within the workings of atomic clocks. To achieve a position accuracy of 10 m requires a clock accuracy of about 3×10^{-8} s, so that after less than an hour an uncorrected GPS will cease to function. The successful operation of the GPS is therefore a practical demonstration of both special and general relativity, in the shape of the equivalence principle. For more details on GPS, see [4.8, 15–17].

In an extraordinary final section to his paper, *Einstein* [4.4] stated that we may *easily* infer, by means of Huyghen's principle, that light rays will be deflected in a gravitational field. He calculates the deflection of light grazing the sun and gets the answer $2MG/(Rc^2) =$ 0.83 arcsec. This is exactly *half* the amount predicted by his later theory of general relativity, and observations show general relativity to be correct. This is the first example of a discrepancy between the equivalence principle and the fully developed theory of general relativity, and invites the following thoughts. The geometry of spacetime is described by the quadratic form $\mathrm{d}s^2$ – see (4.3) above. The equivalence principle enables us to find g_{00} – just *one* component of the metric tensor $g_{\mu\nu}$. All ten components can be found, however (at least in principle), from the Einstein field equations. Test particles (like planets) and light then, in this completely geometric account, simply move along geodesics of the spacetime manifold. In the case of the Sun, Einstein's field equations were solved by Schwarzschild and the bending of light in the sun's filed is found by writing down the null geodesics for this solution – a procedure very different from the one used by *Einstein* in his paper [4.4]!

In the remaining sections of this review, we *improve* the original formulation of the equivalence principle by, firstly, making the definition of acceleration consistent with special relativity and then, in the spirit of general relativity, by finding a spacetime metric which is appropriate to an accelerating frame.

4.4 Relativistic Definition of Acceleration

From his account in [4.3], it is clear that *Einstein* understood (4.2) (and therefore (4.4)) as being an approximation. In fact, in a correction to [4.3], Einstein states [4.18],

> *A letter by Mr Planck induced me to add the following supplementary remark so as to prevent a misunderstanding that could easily arise: in the section Principle of relativity and gravitation a reference system at rest in a temporally constant, homogeneous gravitational field is treated as physically equivalent to a uniformly accelerated, gravitation-free reference system. The concept* uniformly accelerated *needs further clarification.*

Part A | 4.4

If – as in our case – one considers a rectilinear motion (of the system Σ), the acceleration is given by the expression $\mathrm{d}v/\mathrm{d}t$, *where* v *denotes the velocity. According to the kinematics in use up to now,* $\mathrm{d}v/\mathrm{d}t$ *is independent of the state of motion of the (unaccelerated) reference system, so that one might speak directly of (instantaneous) acceleration when the motion in a certain time element is given. According to the kinematics used by us,* $\mathrm{d}v/\mathrm{d}t$ *does depend on the state of motion of the (unaccelerated) reference system. But among all the values of the acceleration that can be so obtained for a certain motion epoch, that one is distinguished which corresponds to a reference system with respect to which the body considered has the velocity* $v = 0$. *It is this value of acceleration which has to remain constant in our 'uniformly accelerated' system. The relation* $v = \gamma t$ *[*γ *is the acceleration]... thus holds only in first approximation; however, this is sufficient, because only terms linear in* t *and* τ*, respectively, have to be taken into account in these considerations.*

An obvious and necessary step toward clarification of this problem is to treat acceleration as a 4-vector in Minkowski spacetime, defined as $\mathrm{d}U^\mu/\mathrm{d}\tau$, where U^μ is the 4-velocity and τ the proper time. It then becomes clear that the (scalar) magnitude of this 4-vector is indeed not constant; it is different in different reference frames. This is shown below.

One occasionally hears it said that special relativity has nothing to say about acceleration, since it is concerned with transformations (Lorentz transformations) which connect one inertial frame with another one. Accelerated frames are noninertial, and therefore beyond the reach of Lorentz transformations. Indeed, the argument would continue, because of the equivalence principle, a passage to an accelerating frame is equivalent to the introduction of a gravitational field, and therefore a passage from special to general relativity. This argument is faulty, however. Gravitational fields are only produced by *heavy* bodies, whereas we are enquiring into the motion of test particles. A discussion of acceleration in the context of special relativity is a discussion that of an accelerating test particle in *Minkowski spacetime*: a test particle will not cause the space to deviate from its Minkowski nature – from flatness.

If the acceleration is in the x direction, we may confine ourselves to the xt plane, so the spacetime position vector in Minkowski space is

$$x^\mu = (x^0, x^1) = (ct, x) \tag{4.5}$$

and the velocity 4-vector is

$$U^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}\tau} = \left(c\frac{\mathrm{d}t}{\mathrm{d}\tau}, \frac{\mathrm{d}x}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}\tau}\right). \tag{4.6}$$

Denoting, as usual

$$\frac{\mathrm{d}t}{\mathrm{d}\tau} = \gamma\,, \quad \frac{\mathrm{d}x}{\mathrm{d}t} = u\,, \quad \gamma = \left(1 - \frac{u^2}{c^2}\right)^{-1/2}, \tag{4.7}$$

then

$$U^\mu = \gamma(c, u)\,, \tag{4.8}$$

and the square of its magnitude is (with metric $-+++$)

$$U^\mu U_\mu = \gamma^2(-c^2 + u^2) = -c^2\,, \tag{4.9}$$

a constant. The acceleration 4-vector A^μ is defined to be

$$A^\mu = \dot{U}^\mu = \frac{\mathrm{d}U^\mu}{\mathrm{d}\tau} = \gamma\frac{\mathrm{d}U^\mu}{\mathrm{d}t} = \gamma\frac{\mathrm{d}}{\mathrm{d}t}(c\gamma, \gamma u)\,. \tag{4.10}$$

Now $u = u(t)$ so γ depends on t. It is straightforward to verify that

$$\frac{\mathrm{d}\gamma}{\mathrm{d}t} = \frac{u}{c^2}\gamma^3 a\,, \quad a = \frac{\mathrm{d}u}{\mathrm{d}t} = \frac{\mathrm{d}^2x}{\mathrm{d}t^2}\,, \tag{4.11}$$

and hence the acceleration 4-vector is

$$A^\mu = \left(\frac{au}{c}\gamma^4, a\gamma^4\right), \tag{4.12}$$

whose square magnitude is

$$A^\mu A_\mu = \alpha^2 = a^2\gamma^6 \tag{4.13}$$

or

$$\alpha = \alpha(u) = \gamma^3 a = \gamma(u)^3\frac{\mathrm{d}u}{\mathrm{d}t}\,. \tag{4.14}$$

It is interesting (and not particularly surprising) that this quantity is not Lorentz invariant. Let us recall that an accelerating particle will possess an instantaneous 4-velocity with respect to an inertial frame. In the instantaneous rest frame $u = 0$; hence

$$\alpha(0) = \frac{\mathrm{d}u}{\mathrm{d}t}\,. \tag{4.15}$$

In a frame moving with respect to this one with a (constant) velocity v, the particle velocity is u', given by

$$u' = \frac{u + v}{1 + \frac{uv}{c^2}} = v$$

and

$$\frac{\mathrm{d}u'}{\mathrm{d}t} = \left(1 - \frac{v^2}{c^2}\right) \frac{\mathrm{d}u}{\mathrm{d}t} = \gamma(v)^{-2} \frac{\mathrm{d}u}{\mathrm{d}t} .$$

Hence

$$\alpha(v) = \gamma(v) \frac{\mathrm{d}u}{\mathrm{d}t} . \tag{4.16}$$

These considerations must be relevant to the query posed by Planck and mentioned by Einstein.

We end this section by noting that the equations (with $A^\mu = \dot{U}^\mu$, $U^\mu = \dot{x}^\mu$)

$$A^\mu U_\mu = 0; A^\mu A_\mu = \alpha^2; U^\mu U_\mu = -c^2 \tag{4.17}$$

can, in the case of *constant* acceleration, be integrated to give [4.9]

$$x = \frac{c^2}{\alpha} \cosh\left(\frac{\alpha\tau}{c}\right) , \quad t = \frac{c}{\alpha} \sinh\left(\frac{\alpha\tau}{c}\right) , \tag{4.18}$$

which lie on the hyperbola

$$x^2 - c^2 t^2 = \frac{c^4}{\alpha^2} . \tag{4.19}$$

In the limit $\alpha\tau/c \ll 1$ these give

$$t = \tau, x = \frac{c^2}{\alpha} + \frac{1}{2}\alpha t^2 = x_0 + \frac{1}{2}\alpha t^2 , \tag{4.20}$$

which describe acceleration in the Newtonian limit.

4.5 Accelerating Frame in Minkowski Spacetime

To begin, consider a simple Newtonian calculation. A particle is projected from a height $y = h$ in the earth's gravitational field, with initial velocity v in the x (horizontal) direction. Newton's laws give

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} = 0 , \quad \frac{\mathrm{d}^2 y}{\mathrm{d}t^2} = -g , \tag{4.21}$$

where g is the acceleration due to gravity. On integration these yield

$$y = h - \frac{g}{2v^2} x^2 , \tag{4.22}$$

a parabola in the xy plane. Our task is to replicate this result (in the nonrelativistic limit) in an account of the motion of particles in an accelerating frame. For this we need to deduce the metric tensor. The motion of particles – and of light – then follows from the geodesic equation.

The relativistic equations for a particle accelerating in the x direction with acceleration a are (4.18) above (with α changed into a)

$$x = \frac{c^2}{a} \cosh\left(\frac{a\tau}{c}\right) , \quad ct = \frac{c^2}{a} \sinh\left(\frac{a\tau}{c}\right) . \tag{4.23}$$

One may describe this motion by a curve given by the vector-valued function of τ,

$$\boldsymbol{P}(\tau) = (ct(\tau), x(\tau)) = \frac{c^2}{a} \left(\sinh \frac{a\tau}{c}, \cosh \frac{a\tau}{c}\right) . \tag{4.24}$$

The *tangent vector* to this curve is

$$\boldsymbol{T}(\tau) = \boldsymbol{P}'(\tau) = c\left(\cosh \frac{a\tau}{c}, \sinh \frac{a\tau}{c}\right) . \tag{4.25}$$

In this (two-dimensional subspace of) Minkowski space these two vectors are orthogonal

$$\boldsymbol{P} \cdot \boldsymbol{T} = \eta_{\mu\nu} P^\mu T^\nu = 0 ,$$

and $\boldsymbol{P}$ is spacelike, and $\boldsymbol{T}$ timelike.

The accelerating observer moves along a timelike worldline (not a geodesic). It is natural for him to use a coordinate system in which he is at rest, in other words, in which his 4-velocity is timelike, as is $\boldsymbol{T}$ above. We may therefore write the four components of the tetrad as

$$\begin{aligned} \left(\frac{1}{c}\boldsymbol{T} =\right) \boldsymbol{e}_0 &= \left(\cosh \frac{a\tau}{c}, \sinh \frac{a\tau}{c}, 0, 0\right) , \\ \left(\frac{a}{c^2}\boldsymbol{P} =\right) \boldsymbol{e}_1 &= \left(\sinh \frac{a\tau}{c}, \cosh \frac{a\tau}{c}, 0, 0\right) , \\ \boldsymbol{e}_2 &= (0, 0, 1, 0) , \\ \boldsymbol{e}_3 &= (0, 0, 0, 1) . \end{aligned} \tag{4.26}$$

In order for the 4-velocity $\boldsymbol{e}_0$ to remain identified with the tangent vector $\boldsymbol{T}$ as the particle moves along its worldline it must be Fermi–Walker transported, as is well known for accelerated motion. We want to describe the metric tensor at a point P *near* the timelike worldline. The nontrivial part of this line, the (ct, x) plane, is

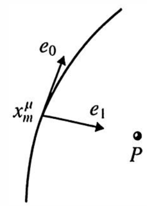


Fig. 4.1 World-line of a particle. P is a nearby point

shown in Fig. 4.1. Suppose the coordinates of a point on the worldline are denoted as x_m^μ. Then we reach the nearby point P by traveling along the *spacelike* vector $\boldsymbol{e}_1$ for a parameter distance x. The so-called Fermi normal coordinates of P, denoted as T, X, are then given by (see, for example, [4.19])

$$cT = \frac{c^2}{a}\sinh\frac{at}{c} + x\sinh\frac{at}{c}$$
$$X = \frac{c^2}{a}\cosh\frac{at}{c} + x\cosh\frac{at}{c}\,, \tag{4.27}$$

where we have replaced τ in (4.24) by t. Then

$$c\mathrm{d}T = \left(1+\frac{ax}{c^2}\right)\cosh\frac{at}{c}c\mathrm{d}t + \sinh\frac{at}{c}\mathrm{d}x$$
$$\mathrm{d}X = \left(1+\frac{ax}{c^2}\right)\sinh\frac{at}{c}c\mathrm{d}t + \cosh\frac{at}{c}\mathrm{d}x\,,$$

and

$$-c^2\,\mathrm{d}T^2 + \mathrm{d}X^2 = -\left(1+\frac{ax}{c^2}\right)^2 c^2\,\mathrm{d}t^2 + \mathrm{d}x^2\,. \tag{4.28}$$

The metric tensor at P is then

$$g_{\mu\nu} = \begin{pmatrix} -\left(1+\frac{ax}{c^2}\right)^2 & 0 \\ 0 & 1 \end{pmatrix}. \tag{4.29}$$

This is the metric for an accelerating particle in Minkowski space. Other derivations of this metric may be found in [4.2, 20, 21]. It is straightforward to check that (4.29) describes a *flat* space, as of course it should. It is also interesting to check that, in the nonrelativistic limit, it describes a parabolic path, (4.22), for the motion of particles in the plane. To see this, consider the $(2+1)$ space (x, y, t), with acceleration in the y direction (vertical). The metric tensor is

$$g_{\mu\nu} = \begin{pmatrix} -\left(1+\frac{ay}{c^2}\right)^2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{4.30}$$

With $ay/c^2 \ll 1$ we may put

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$$

and to lowest order

$$h_{00} = -\frac{2ay}{c^2}\,, \quad \text{other} \quad h_{\mu\nu} = 0\,.$$

The connection coefficients are, to lowest order, given by

$$\Gamma^\lambda_{\mu\nu} = \frac{1}{2}\eta^{\lambda\kappa}\left(h_{\kappa\mu,\nu} + h_{\kappa\nu,\mu} - h_{\mu\nu,\kappa}\right),$$

yielding

$$\Gamma^2_{00} = \frac{a}{c^2}\,, \quad \text{other} \quad \Gamma^\lambda_{\mu\nu} = 0\,.$$

The geodesic equation

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}s^2} + \Gamma^\mu_{\kappa\lambda}\frac{\mathrm{d}x^\kappa}{\mathrm{d}s}\frac{\mathrm{d}x^\lambda}{\mathrm{d}s} = 0$$

then gives, for $\mu = 1, 2$,

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = 0\,, \quad \frac{\mathrm{d}^2y}{\mathrm{d}t^2} = -a\,,$$

exactly as in (4.21) above, with solution (4.22) – parabolic motion – thereby confirming the equivalence principle for free fall with metric (4.30).

Let us finally consider the propagation of light in an accelerating frame. Light obeys $\mathrm{d}s^2 = 0$, and the metric (4.30) then gives, to leading order in ay/c^2,

$$\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2 - c^2\left(1+\frac{2ay}{c^2}\right) = 0\,.$$

This is solved by $x = ct$, $y = \frac{1}{2a}t^2$, i.e. $y = \frac{a}{2c^2}x^2$ – the equation of a *parabola* (to lowest order), exactly as deduced from very simple arguments. Hence light, as viewed in an accelerating frame, does not travel in a straight line in a flat space – just as, of course, it does not in a curved spacetime.

4.6 Concluding Remarks

There is one more topic often raised in connection with the equivalence principle, which is the fact that a charged particle will emit electromagnetic radiation when accelerated, so the equivalence principle would imply that it should do so also in a static gravitational field, but that is not observed. Does this count as a violation of the equivalence principle? This question has been around for a long time; it has proved somewhat contentious and attempts to resolve it involve rather nontrivial considerations. This is not the place to attempt any kind of summary, but highly readable recent accounts may be found in [4.22, 23].

We may conclude in the following way. Einstein's insight, that a gravitational field is locally equivalent to an accelerating frame, was a major step toward his formulation of general relativity in 1916. General relativity, however, contained one crucial ingredient, the curvature of spacetime, which is missing from the equivalence principle. It is this notion which gives general relativity its distinctive character, and the (slowly increasing number of) tests which verify this theory, for example the Gravity Probe B experiment, all rely on spacetime curvature. Minkowski spacetime, even when viewed from an accelerating frame, is flat, so the equivalence principle cannot be taken seriously as a theory of gravity. Its real concern is the similarity of inertial forces and gravitational ones. These are different physical phenomena, but the equivalence principle dramatically highlights a similarity between them. The relation between gravity and inertia is a subject that needs to be understood more deeply.

References

4.1 A. Pais: *Subtle is the Lord* (Oxford Univ. Press, New York 1982)
4.2 C.W. Misner, K.S. Thorne, J.A. Wheeler: *Gravitation* (Freeman, San Francisco 1973)
4.3 A. Einstein: Jahr. Rad. Elektr. 4, 411 (1907), translated in The collected papers of Albert Einstein, Vol. 2 (Princeton Univ. Press, Princeton 1989) p. 252
4.4 A. Einstein: Ann. der Physik, 35, 898 (1911), translated in H.A. Lorentz, A. Einstein, H. Minkowski, H. Weyl (eds.): The Principle of Relativity (Dover, New York 1952), p. 99, and in The Collected Papers of Albert Einstein, Vol. 3 (Princeton Univ. Press, Princeton 1993), p. 379
4.5 A. Einstein: Ann. der Physik 49, 769 (1916), translated in H.A. Lorentz, A. Einstein, H. Minkowski, H. Weyl (eds.): The Principle of Relativity (Dover, New York 1952), p. 111, Dover Publications, New York (1952), and in The Collected Papers of Albert Einstein, Vol. 6 (Princeton Univ. Press, Princeton 1997), p. 146. See also A. Einstein: The Meaning of Relativity (Methuen, London 1967)
4.6 J.L. Synge: *Relativity: The General Theory* (North-Holland, Amsterdam 1964)
4.7 Y. Su, B.R. Heckel, E.G. Adelberger, J.H. Grundlach, M. Harris, G.L. Smith, H.E. Swanson: New tests of the universality of free fall, Phys. Rev. D **50**, 3614 (1994)
4.8 J.B. Hartle: *Gravity: An Introduction to Einstein's General Relativity* (Addison-Wesley, San Francisco 2009)
4.9 L. Ryder: *Introduction to General Relativity* (Cambridge Univ. Press, Cambridge 2009)
4.10 M.M. Nieto, R.J. Hughes, T. Goldman: Actually, Eötvös did publish his results in 1910, it's just that no one knows about it, Am. J. Phys. **57**(5), 397 (1989)
4.11 H. Müller, A. Peters, S. Chu: A precision measurement of the gravitational redshift by the interference of matter waves, Nature **463**, 926 (2010)
4.12 P. Wolf, L. Blanchet, C.J. Borde, S. Reynard, C. Salmon, C. Cohen-Tannoudji: Atom gravimeters and gravitational redshift, Nature **467**, E1 (2010)
4.13 H. Müller, A. Peters, S. Chu: Atom gravimeters and gravitational redshift, Nature **467**, E2 (2010)
4.14 D. Giulini: Equivalence principle, quantum mechanics and atom-interferometric tests. In: *Quantum Field Theory and Gravity: Conceptual and Mathematical Advances in the Search for a Unified Framework*, ed. by F. Finster, O. Müller, M. Nardmann, J. Tolksdorf, E. Zeidler (Birkhäuser, Boston 2012)
4.15 B. Schutz: *Gravity from the Ground Up* (Cambridge Univ. Press, Cambridge 2009)
4.16 N. Ashby: Relativity and the Global Positioning System, Phys. Today **55**(5), 41 (2002)
4.17 N. Ashby: Relativity and the Global Positioning System, Living Rev. Relativity 6, 1 (2003), available online at www.livingreviews.org/lrr-2003-1
4.18 A. Einstein: Jahrb. Rad. Elektr. 5, 98 (1908), translated in The Collected Papers of Albert Einstein, Vol. 2 (Princeton Univ. Press, Princeton 1989) p. 316
4.19 F.K. Manasse, C.W. Misner: Fermi normal coordinates and some basic concepts in differential geometry, J. Math. Phys. **4**, 735 (1963)
4.20 F.W. Hehl, Y.N. Obukhov: *Foundations of Classical Electrodynamics* (Birkhäuser, Boston 2003), Section E.4.8

4.21 S.N. Lyle: *Self-force and Inertia: Old Light on New Ideas* (Springer, Berlin 2010)
4.22 S.N. Lyle: *Uniformly Accelerating Charged Particles. A Threat to the Equivalence Principle* (Springer, Heidelberg 2008)
4.23 S.N. Lyle: Electromagnetic radiation and the coming of age of the equivalence principle, Problems and Developments of Classical Electrodynamics, Proc. 475th WE-Heraus-Seminar Bad Honnef (2011)

5. The Geometry of Newton's and Einstein's Theories

Graham S. Hall

The aim of this paper is to present a simple, brief, mathematical discussion of the interplay between geometry and physics in the theories of Newton and Einstein. The reader will be assumed to have some familiarity with classical Newtonian theory, the ideas of special and general relativity theory (and differential geometry), and the axiomatic formulation of Euclidean geometry. An attempt will be made to describe the relationship between Galileo's law of inertia (Newton's first law) and Euclid's geometry, which is based on the idea of Newtonian absolute time. Newton's second law and classical gravitation theory will then be introduced through the elegant idea of Cartan and his space-time connection and space metric. This space metric will then be used to introduce Minkowski's metric in special relativity and its subsequent generalization, by Einstein, to incorporate relativistic gravitational theory. The role of the principles of equivalence and covariance will also be discussed. Finally, a brief discussion of cosmology will be given. Stress will be laid on the (geometrical) concepts involved rather than the details of the mathematics, in so far as this is possible.

5.1 Guide to Chapter

The object of this paper is to give an elementary conceptual introduction to the geometry of classical Newtonian theory and Einstein's special and general theories of relativity. It is assumed that the reader has a knowledge of elementary geometry including that of Euclid and also some differential geometry, but it is also intended that nothing of a really technical nature will be involved in the paper. It begins with a brief discussion of Euclidean geometry and its relationship to the so-called non-Euclidean geometry of Lobachevski and Bolyai, the axiomatization of Euclidean geometry by Hilbert and the extension of geometry to the concept of the metric and to what would now be called manifold theory, by Riemann. Next, the formalism of Newton's theory and the ideas of Newtonian gravitational theory are introduced together with the relationship between Newton's ideas and Euclidean geometry. This is followed, in Sect. 5.4, by a statement of the Newtonian principle of equivalence in gravitation theory and a discussion of Cartan's development of Newtonian gravity by his introduction of a connection on Newtonian space and time, whose geodesics are the paths of freely falling particles under the action of a gravitational force. Although Cartan's ideas came some years after Einstein's work on general relativity, it so clearly expresses Einstein's ideas, and in a more elementary way, that its inclusion can be justified as an introduction to certain techniques of general relativity theory. In Sect. 5.5, special relativity is introduced and compared with classical space-time theory. Here, the introduction of the null cone structure and the Minkowski metric is discussed together with its place in the development of

general relativity theory. In Sect. 5.6, a brief discussion of the principle of covariance in physics and its relation to the distinction between *absolute* and *dynamical* variables is presented, again in preparation for the work on general relativity. In Sect. 5.7, general relativity theory is introduced and Einstein's field equations are briefly explained. No attempt is made to enter into the details of these equations. Rather, this section will be simply a discussion regarding the justification for and the nature of them and, in particular, their geometric content. In the final section a brief geometrical introduction to cosmology is presented with emphasis on the (physical) symmetry assumptions made in such a study.

5.2 Geometry

Over 2000 years have elapsed between Euclid's *elements* and the modern developments in axiomatic geometry initiated by *David Hilbert* [5.1]. In that period, Euclidean geometry was essentially believed to apply to *space* and, as such, was the backbone of the physical sciences and, in a sense, a branch of applied mathematics. Its origins, as the name suggests, were in land measurement, and it was the subject of experimental tests to verify its accuracy (although, in fact, such tests could obviously only refute its accuracy and not confirm it). Since it was seen as a *visual* science, theoretical work in geometry was hampered by imprecisions in its (intuitive) foundations. Euclid, in writing the *elements*, laid the foundations not only of geometry but also of a (limited type of) axiomatic method. He started with certain unquestioned assumptions and used them (sometimes not entirely faithfully) to derive the *theorems* of this subject. As a body of knowledge it stood supreme for over 2000 years with the only dispute of a fundamental nature being that over whether the *parallel postulate*, introduced as an *axiom* by Euclid (albeit in a different form than usually understood now), could be derived from his other axioms and hence reduced to a theorem. This dispute was settled by the independent work of the Russian, Lobachevski, and the Hungarian, Bolyai, and their discovery of what is called *non-Euclidean* or *hyperbolic* geometry. (Lobachevski was the first to announce his work when he presented it to the physical-mathematical division of the University of Kazan in 1826 and published it in the *Kazan Messenger* in 1829. Bolyai first published his work as an appendix to his father's mathematics book in 1831. An earlier announcement of this geometry was claimed by Gauss. An excellent history of such things can be found in the books by *Bonola* [5.2] and *Meschkowski* [5.3].) The geometry of Lobachevski and Bolyai satisfied all the axioms of Euclid except his parallel postulate and thus showed that this latter postulate could not be derived as a theorem from the remaining axioms. However, it is not clear that this non-Euclidean geometry was received with anything more than theoretical interest and Euclid's geometry was still the means of navigating space. In the middle of the nineteenth century geometry took a different turn through the work of *Bernhard Riemann* [5.4] and the initial ideas concerning manifolds and metrics on them and the *curvature* that they generated. Riemann's geometry essentially reduced Euclidean and non-Euclidean geometry to the status of *special cases*, his concept of (a possibly varying) curvature giving greater flexibility to the choice of model. His work not only inspired many new developments in the analytical nature of geometry but was fundamental in Einstein's general theory of relativity. This will be discussed further later.

At the beginning of the twentieth century much energy in mathematical research was expended in exploring the foundations of mathematics after the latter had been damaged by the Russell paradox (for a review of this history, see [5.5]). This led Hilbert to fortify Euclidean geometry by establishing a modern axiomatic foundation for it. This had an impact in two directions; firstly, it directed attention to the concept of a formal axiomatic system for geometry where the objects (points, lines, planes) to which the axioms applied were undefined (primitive) elements and could be interpreted in any desired way which was consistent with the axioms. This removed the problem over arguments about what a *point*, a *line*, or a *plane* was in Euclid's scheme since, in Hilbert's formulation, all that mattered was whether members of the (nonempty) sets $\mathcal{P}$ (of points), $\mathcal{L}$ (of lines), and Π (of planes) together satisfied certain imposed conditions such as an *incidence* relation $\mathcal{I} \subset \mathcal{P} \times \mathcal{L}$ on the set $\mathcal{P} \times \mathcal{L}$ for which $(p, L) \in \mathcal{I}$ $(p \in \mathcal{P}, L \in \mathcal{L})$ could be (but need not be) interpreted as saying that the point p was *incident* with the line L. Thus Hilbert placed geometry in the realms of pure mathematics where it could be inspected logically, rather than a form of space science bedevilled

by pointless controversy. Secondly, Hilbert's work laid down the foundations for the various axiomatic systems that dominate modern mathematics. Modern formulations of Hilbert's work can be found in [5.6–9].

To understand Hilbert's ideas it is sufficient to study them in two-dimensional form where only the sets $\mathcal{P}$ and $\mathcal{L}$ and the incidence relation given above are relevant. Hilbert's axioms controlling them come in five groups which, in a modern formulation and ordering, are:

i) Axioms of incidence
ii) Axioms of betweenness
iii) Axioms of congruence
iv) The completeness axiom
v) The Euclidean parallel axiom.

These axioms, with the possible exception of (iv), are beautifully intuitive and collectively *categorical* in the sense that there is only one model (up to an obvious geometrical isomorphism) satisfying them, namely, the Euclidean plane $\mathbb{R}^2$ with its usual straight line structure. These axioms leave sufficient room for a unit of length and angular measure to be chosen. Further, by using the obvious definition of *parallel* and replacing the Euclidean parallel axiom (v) above by (v′) *for any given $p \in \mathcal{P}$ and $L \in \mathcal{L}$ with (p, L) not in $\mathcal{I}$ there exist at least two distinct members of $\mathcal{L}$ with which p is incident and which are parallel to L*, one achieves a categorical axiomatization of the Lobachevski–Bolyai geometry. Granted the axioms (i)–(iv) these two geometries are the only options since the existence of one such *parallel* line is guaranteed by axioms (i)–(iv) (and, in fact, it then follows that if more than one such line exists then infinitely many do, and all models thus arising are geometrically isomorphic to the geometry of Lobachevski and Bolyai). Since the work of Hilbert, many other axiom systems for Euclidean geometry have appeared, including one which explores the metrical properties of this geometry [5.6] and is based on the idea of a metric space and, again, is easily modified to categorize Lobachevski–Bolyai geometry. Other approaches have been taken, some of which are based on the idea of symmetry, others purely on arithmetic, and others on the idea of testing how far one can go without the (somewhat different and *ungeometrical*) completeness axiom and introducing, for example, a circle–circle intersection property [5.9]. The axiomatization given here may be extended to the usual three-dimensional Euclidean space $\mathbb{R}^3$ by the introduction of the set Π of planes together with some appropriate modifications to the axioms, which are also categorical. It is this three-dimensional geometrical structure that will be important in what is to follow.

What is the role of geometry in physics? Is it a mere convenience for the description of phenomena, set in stone, whilst the laws of physics are modified to fit it, or is it a part of physics, flexible enough to be modified by the presence of bodies etc., and forming a coalition with physics, allowing its canonical structures (lines etc.) to be the descriptors of fundamental physical laws? (For different viewpoints on this see [5.10] and [5.11].) If space is Euclidean, as Kant would have us believe is the only real possibility, then it seems that we would want to say something along the following lines. Three-dimensional Euclidean space (regarded as $\mathbb{R}^3$ with Hilbert's axioms suitably imposed on it, with lines and planes represented by the usual linear relationships between coordinates and including a measure of length and angle consistent with Pythagoras' theorem) can now be regarded as a three-dimensional real vector space and admits a group of transformations, the Euclidean group, which is a group of global maps $\mathbb{R}^3 \to \mathbb{R}^3$ consisting of (combinations of) linear orthogonal transformations and translations. With this given definition of length and angle in $\mathbb{R}^3$, such maps preserve angles and lengths, and hence *shape*. Suppose one speaks (loosely) of a body being moved from one point in space to another without change of shape. If this means anything at all it presumably means that it may be moved pointwise by means of one of these maps. Thus one has the concept of a *rigid body*. In this way, one has made the assumption that $\mathbb{R}^3$ has a geometry which, in some sense, *reacts* with its contents rather than just being a convenient arena for the description of events. Then one can devise a physical measuring rod which may be constructed so as to be brought into coincidence with part of any of the straight lines defined by the geometry and can be used as a means of consistently transporting a measure of length around space which coincides everywhere with the length prescribed geometrically on $\mathbb{R}^3$. Similar devices may be constructed for the measurement of angles. Another way of expressing this is to say that the usual methods of physical measurement on our space are found to be consistent with a Euclidean structure on that space.

5.3 Newtonian Mechanics I

In this section the coordinatization of the universe of events E and the introduction of Galileo's law of inertia will be undertaken. It is based, in part, on the work in [5.12, 13]. In Newtonian theory it is customary to view the universe of events E as controlled by Newton's *absolute time*. Absolute time is taken as a function $T : E \to \mathbb{R}$ and, in tandem with it, E is equipped with the usual (sufficient) collection of *good clocks* such that any good clock used at any event $e \in E$ would measure the time $T(e)$. This is, of course, independent of the clock's motion; in fact motion has not yet been defined. The *instantaneous* or *simultaneity* spaces of constant absolute time, $S_t \equiv T^{-1}\{t\}$ for each $t \in \mathbb{R}$, are each assumed to be of the same cardinality as $\mathbb{R}^3$ and to carry a three-dimensional Euclidean structure as described in the previous section. One may then choose, in infinitely many ways, Cartesian coordinates x, y, z on each set S_t such that the latter's lines and planes within this Euclidean structure are described, as usual, by linear relations between the coordinates x, y, z, the units of the latter equalling the unit of length chosen. Any two of these coordinate systems are then connected by a member of the Euclidean group mentioned earlier. An *observer* will then be thought of as a coordinate system consisting of a choice of such coordinates in each S_t together with the absolute time coordinate t, so that each $e \in E$ has coordinates of the form (x, y, z, t). (In preparation for what is to come, it is noted that, by their very definition, each set of events S_t is independent of the observer and that this fact is a consequence of the definition of absolute time.)

However, no control has been placed on the way the *spaces* S_t are related to each other for different values of t. For example, if a particle P moves through E one might wish, intuitively, to express the *continuity* of its motion with respect to some observer by stipulating that its path, as a function of (absolute) time t with t drawn from some open interval $I \subset \mathbb{R}$ and given by $t \to (x(t), y(t), z(t))$, yields three continuous functions x, y, and z from I to $\mathbb{R}$ when the usual topologies are placed on $\mathbb{R}$ and I.

One way of relating the spaces S_t can now be described. First, one recalls that Newtonian theory essentially claims to recognize the existence of a situation free of *real* forces (gravitational, electromagnetic, etc.) and the existence of *free* particles. Then one assumes the existence of a collection of observers called *inertial observers* for each of whom the Cartesian coordinates described above on each S_t may be chosen so that the path of any free particle is given by a map $t \to (x(t), y(t), z(t))$ with $x(t), y(t)$ and $z(t)$ *linear* functions of t. It is also assumed that any such triple of linear functions is the (potential) path of some free particle. Thus, each inertial observer gives rise to a map $E \to \mathbb{R}^4$ by attaching to an event $e \in E$ the 4-tuple (x, y, z, t) where $t = T(e)$ and (x, y, z) are the (projected) coordinates of $e \in S(t)$ and this provides a global chart for a four-dimensional manifold structure on E. The path of any free particle is then a (smooth) linear map $I \to E$. Any other inertial observer similarly gives rise to a global chart with coordinates (x', y', z', t), and it is assumed that the collection of global coordinate systems obtained in this way gives rise to a smooth atlas on E. Then the coordinate transformation between any two of them is one of the usual Galilean transformations. For any given inertial observer O a particle whose path is given by $t \to (x(t), y(t), z(t))$ with x, y and z constant (hence linear) functions is then said to be *at rest* in O and gives rise, in an obvious way, to a *fixed point* in O.

Finally, for a fixed inertial observer O and for any fixed $t_0 \in \mathbb{R}$ let $f_t : S(t) \to S(t_0)$ be the identity map in (the restrictions to $S(t_0)$ and $S(t)$ of) the coordinates of O. This map preserves the distance and angle constructions placed initially on $S(t_0)$ and $S(t)$ for each t and each inertial observer O. Then the path of any free particle can be projected, using this map, onto $S(t_0)$, giving the more usual particle path $t \to (x(t), y(t), z(t))$ in the fixed Euclidean space $S(t_0)$ for linear functions $x(t), y(t)$ and $z(t)$ of the parameter t so that, effectively, the fixed particles are acting as coordinate reference points. Using the projection map f_t, one has a measure of spatial distance in O's frame, between any two events in E by projecting each of these events onto $S(t_0)$ and computing the distance between them there in O. In this sense the link between the Euclidean geometry of space and Galileo's law of inertia (Newton's first law) is displayed. One can think of the triple $(E, T, \mathbb{R})$ as a bundle with smooth projection map T which, although diffeomorphic to $\mathbb{R}^4$ and possessing a product structure for each inertial observer, has no *natural* preferred product structure (that is, no preferred inertial observer). If, on the other hand, one introduces a strictly absolute space (*Aristotelian spacetime*) with its preferred observer *at rest* (as opposed to a collection of absolute spaces, one for each inertial observer) or classical Maxwell electrodynamics with its preferred *ether frame* (an inertial frame in which the (rigid) ether is at rest and hence in which the speed of light is independent of its direction)

then a natural product structure is singled out. On the philosophical problems of any form of absolute space the critique by *Mach* [5.14] is illuminating.

In classical theory, it was also an experimental result that, *in vacuo*, light travels in *straight lines* in the sense that the projected paths of light photons onto $S(t_0)$, as above, also satisfy linear equations in t. Thus, both Galilean–Newtonian mechanics and Maxwell's electromagnetic theory determine, in this sense, the same family of lines in each $S(t)$.

5.4 Newtonian Mechanics II

When considering the motion of a particle P in Newtonian theory under the action of a force, it is usual to attribute to P a *mass* m which is assumed constant. When Newton's second law is introduced, m assumes the role of a coupling constant and the force vector acting on P in some inertial frame is then the product of m and the acceleration vector of P in that frame. In this sense, m is the inertial mass of P, that is, a measure of P's resistance (inertia) to being accelerated. If this were simply the definition of force, little would have been achieved. The real content of Newton's second law emerges when this force is independently specified, as it is, for example, during a (Newtonian) gravitational interaction using the inverse square law. To see this in more detail consider two particles A and B attracting each other gravitationally and upon which no other forces act. Suppose that in some inertial frame and at absolute time t, A and B have position vectors $\boldsymbol{r}_A(t)$ and $\boldsymbol{r}_B(t)$, respectively. Suppose now that one attaches three (constant) mass parameters to each particle; its *active gravitational mass* (its ability to attract a particle, gravitationally) denoted by a subscript AG, its *passive gravitational mass* (its susceptibility to being gravitationally attracted) denoted by a subscript PG), and its *inertial mass* (its resistance to being accelerated with respect to an inertial frame) denoted by the subscript I). Then with M_{AG}, M_{PG}, and M_{I} denoting these respective mass parameters for particle A and similarly, using the symbol m for B, Newton's third law together with his second law augmented by his inverse square law gravitational hypothesis, give

$$\begin{aligned} \frac{GM_{\text{AG}}m_{\text{PG}}}{r^2} &= \frac{GM_{\text{PG}}m_{\text{AG}}}{r^2}\,, \\ m_{\text{I}}\ddot{\boldsymbol{r}}_B &= \frac{GM_{\text{AG}}m_{\text{PG}}}{r^2}\,, \end{aligned} \tag{5.1}$$

where $r(t)$ is the distance between A and B at time t, G is the Newtonian gravitational constant, and a dot denotes $\mathrm{d}/\mathrm{d}t$. The first of these shows that $M_{\text{AG}}/(M_{\text{PG}}) = m_{\text{AG}}/m_{\text{PG}}$ and so, by assuming that this is true for all particles and at all points of space and time, one may choose units of active and passive gravitational mass so that the active and passive gravitational masses are the same for every particle. Calling this common mass parameter the *gravitational mass* and denoting it M_G (and m_G) the second equation in (5.1) reads $m_{\text{I}}/(m_G\ddot{\boldsymbol{r}}_B) = (GM_G)/r^2$. Since the right-hand side of this equation depends only on the body A and the position of B, one makes the assumption that at each event in the universe E, for a fixed gravitational field, the quantity $m_{\text{I}}/(m_G\ddot{\boldsymbol{r}}_B)$ is independent of B. The experiments of Eötvös, Dicke, and others (see, for example, [5.15, Chap. 38]) provide strong evidence that the (gravitational) acceleration of B, $\ddot{\boldsymbol{r}}_B$, is separately fixed under such circumstances and hence one makes the assumption that m_{I}/m_G is the same for all particles at all events. Units may then be chosen so that $m_G = m_{\text{I}} (= m)$, giving a single *mass* parameter m for every particle. In this sense one has the classical result that the equality of the inertial and gravitational masses is equivalent to the fact that, in a given gravitational field, a well-defined *gravitational acceleration* exists. This is one form of the *Newtonian principle of equivalence*. (The fact that the behavior of a simple pendulum of fixed length at a fixed place on the Earth's surface is independent of the mass of the pendulum bob is another consequence of the equality of the inertial and gravitational masses of the bob.) Such a result arises as a consequence of Newton's laws and reveals the indiscriminate nature of the gravitational force, caring nothing for the mass or make-up of the body on which it acts, but imparting to it a certain acceleration with respect to an inertial frame and which is independent of its mass. This result thus appears as a theorem in Newtonian theory (and was well known to Galileo and Newton).

Conventional Newtonian theory thus declares the path of a particle in a given gravitational field to be (at least locally) determined if its velocity is specified at some event in E, since then the fixing of the acceleration and Newton's second law combine to give a second-order differential equation complete with ini-

tial conditions. This leads to an alternative viewpoint on Newton's equations. Since the force-free situation in Newtonian theory has already been described geometrically by a Euclidean structure in each of the spaces $S(t)$ and in keeping with the discussion of general relativity to come, one asks if some other geometrical structure can be put on E which for some gravitational field in E, in some sense, itself yields a family of curves in E such that for each $e \in E$ and potential particle velocity $\boldsymbol{v}$ at e, exactly one such curve passes through e with tangent vector $\boldsymbol{v}$ at e, and which would represent the path of a particle under the influence of the gravitational field *without the need of forces*. Then *pure* gravitational fields would have been *geometrized*, that is, no force is explicitly given but rather a geometry determined, say, by the sources of the gravitational field and for which certain canonical paths will then decide the motion of particles under the influence of such sources.

Such a program was initiated by *Cartan* ([5.16], see also [5.12, 13, 15]) some years after Einstein published his general theory of relativity. This involved introducing a gravitational field and starting, say, with a certain special observer O. One of the consequences of this work was that it revealed the difficulties with the concept of an inertial observer. Consider the set E described in Sect. 5.3 and let O be this observer with coordinates $(x, y, z, t) \equiv (x^\alpha, t)$, with Greek letters taking the values 1, 2, and 3 (and $x^1 = x$, $x^2 = y$, $x^3 = z$). When the gravitational field is introduced into E and described in O's frame by a gravitational potential $\Phi : E \to \mathbb{R}$ let a tentative definition that O is an inertial observer be that O is such that the path $t \to (x^\alpha(t), t)$ of a *freely falling particle* P in this gravitational field and in O's coordinate system satisfies

$$\frac{\mathrm{d}^2 x^\alpha}{\mathrm{d}t^2} = \frac{\partial \Phi}{\partial x^\alpha} . \tag{5.2}$$

It can then be noticed that P's path, with absolute time as parameter and $x^0 = t$, satisfies (using the usual Einstein summation convention)

$$\frac{\mathrm{d}^2 x^a}{\mathrm{d}t^2} = \Gamma^a_{bc} \frac{\mathrm{d}x^b}{\mathrm{d}t} \frac{\mathrm{d}x^c}{\mathrm{d}t} , \tag{5.3}$$

where Latin letters take the values 1, 2, 3, 4, and Γ^a_{bc} represent certain functions defined on E and determined entirely by the gradient of the potential Φ as

$$\Gamma^\alpha_{00} = \frac{\partial \Phi}{\partial x^\alpha} , \tag{5.4}$$

with all other $\Gamma^a_{bc} = 0$. Thus, $\Gamma^a_{bc} = \Gamma^a_{cb}$. The general idea is to use this information to define a *symmetric connection* Γ on E by stipulating that the functions Γ^a_{bc} are the connection coefficients of Γ in this coordinate system (and thus they transform according to the standard law for such coefficients under a change of coordinates in E). Then (5.3) reveals that the motion of P is a geodesic of Γ with t as an affine parameter. The standard result is noted at this point that such a geodesic is uniquely determined by a point through which it passes together with its *tangent direction* spanned by $\frac{\mathrm{d}x^a}{\mathrm{d}t} = (\boldsymbol{u}, 1)$ at that point and is thus fixed by that point and the particle velocity $\boldsymbol{u}$ at that point.

Now suppose that O' is another observer describing the (geodesic) motion of P. Then O''s coordinates (x', y', z', t), assumed to be in the atlas placed on E and with x', y', z' Cartesian coordinates in each section $S(t)$, are related to those of O by a space rotation, represented by an orthogonal 3×3 matrix $A(t)$ and a translation, represented by a three-vector $\boldsymbol{a}(t)$ where each entry in $A(t)$ and each component of $\boldsymbol{a}(t)$ are smooth and so

$$\begin{aligned} x'^\alpha &= A^\alpha{}_\beta(t) x^\beta + a^\alpha(t) , \\ t' &= t . \end{aligned} \tag{5.5}$$

One now asks how O' would view the motion of P not forgetting that he must also regard it as a geodesic with respect to Γ with affine parameter t. Naively, one might say that the forces that O' regards as acting on P consist of pure gravitational forces similar to those described by the right-hand side of (5.2) together with *inertial type* forces reflecting, in some sense, O''s motion with respect to O. Suppose, in an attempt to be able to call O' *inertial* also, one assumes that these latter (inertial) forces are absent and insists that O''s description of the motion of P is as in (5.2), so that $(\mathrm{d}^2 x'^\alpha)/\mathrm{d}t^2 = \partial\Phi'/\partial x'^\alpha$, where $\Phi' : E \to \mathbb{R}$ is the gravitational potential in O'. Since the motion of P is a geodesic of Γ with t as an affine parameter, (5.3) and (5.4) must hold with primes in the appropriate places. Then, on transforming the functions Γ^a_{bc} under (5.5) in the usual way for such connection coefficients, one finds that the stipulated conditions are satisfied if and only if A is a *constant* (orthogonal) matrix and (up to an arbitrary constant) $\Phi' = \Phi + \ddot{a}^\alpha x'^\alpha$ for a time-dependent vector $\boldsymbol{a}(t)$, and where a dot denotes $\mathrm{d}/\mathrm{d}t$ (see, e.g., [5.12, 15]). The conclusion is that the space coordinates of O and O' are related by a rotation which is the same in each space $S(t)$ together with a *time-dependent* translation. An important conclusion is that the potential function is *not* well defined on E since Φ and Φ' are differ-

ent, in general (and by more than an arbitrary constant). Thus Φ depends on a coordinate system for its value and cannot be measured in a coordinate-free way. One may view the inability to determine the potential function as the inability to determine the term $\ddot{a}^\alpha x'^\alpha$ and hence the inability to determine the difference between the connection components Γ'^a_{bc} and Γ^a_{bc}. The difference between these components then gives rise to a *flat* connection on E since its components can be transformed to zero. Thus the above inability to determine the potential function now reappears as the inability to determine this flat connection. (One may object to this by saying that since O is assumed *inertial*, O' is not inertial unless $\dot{a}(t)$ is identically zero, from (5.5), and then $\Phi' = \Phi$. But the point being made is that one cannot distinguish between O and O' given that one wishes to retain the features that the gravitational potential is a real-valued function on E and that (5.2) holds for O and O'. This reflects the fact that a linear acceleration represented by $\boldsymbol{a}(t)$ with $\dot{\boldsymbol{a}}(t)$ not the zero function would not change the form of (5.2) in the passage from O to O' and that the inertial forces so generated are, like gravitational forces, indiscriminate in their accelerative effect on particles (and simulate a *homogeneous* gravitational field).) Of course, if one could (using the freedom of the choice of arbitrary constant in Φ and Φ') say that Φ and Φ' each *tend to zero as one moved out to large distances* (as, for example, would be the case if P were under the influence of an isolated *island universe* of gravitational sources) then one does achieve the condition that $\dot{\boldsymbol{a}}(t)$ is a constant function, but modern cosmological observations suggest that such a physical situation is not the case.

In summary, one can say that, given that a certain observer O in the usual force-free case of Sect. 5.3 is designated *inertial*, all other inertial observers are found by applying the Galilean group of transformations to O, whereas if a gravitational field is introduced and O is an inertial observer such that (5.2) holds, then for O' to get a similar equation for the particle P's equation of motion the transformations linking O and O' constitute a larger group and are given by (5.5). In this sense, in a gravitational field, the concept of an inertial observer is lost and this because of the inability to distinguish inertial forces (brought about in this case through the term $\boldsymbol{a}(t)$) from gravitational ones. To put it another way, having combined the *inertial* forces due to $\boldsymbol{a}(t)$ with those from Φ in some coordinate system, they cannot be unambiguously recovered from it. In short, all that can be determined are the geodesics (5.3) and hence the connection Γ, and so the gravitational field has been *geometrized* by imposing the connection Γ in E. (In a strictly Newtonian theory such a separation of gravitational and inertial forces, and hence inertial frames, would be assumed possible.)

The elementary geometrical properties of the symmetric connection Γ are easily explored. It is clear from the definition (5.4) in the coordinate system of O that the global vector fields $X = \partial/\partial x$, $Y = \partial/\partial y$ and $Z = \partial/\partial z$, the global 1-form $\mathrm{d}t$ and the global, second-order, symmetric, everywhere rank three tensor $h \equiv X \otimes X + Y \otimes Y + Z \otimes Z$ (with components diag(1, 1, 1)) are Γ-covariantly constant. Thus the three-dimensional hypersurfaces of constant t (space sections) are invariant under parallel transport in the sense that any vector $v \in T_eE$, where T_eE denotes the tangent space to E at e, which is tangent to these hypersurfaces, remains so after parallel transport along any closed curve at e. In fact, v will return exactly to v under such circumstances. It follows that Γ induces a connection in these hypersurfaces, which is flat and that one may take h as a flat metric in these hypersurfaces compatible with the induced connection. Thus the *Euclidean* structure of the space sections is recovered. From (5.4) one can compute the curvature tensor Riem associated with Γ in the coordinates of O to get [5.15]

$$R^\alpha{}_{0\beta 0} = \frac{\partial^2 \Phi}{\partial x^\alpha} \partial x^\beta \,, \tag{5.6}$$

with all other components of Riem not contained in (5.6) zero. From this it can be concluded that Γ is not a *metric connection* unless it is a flat connection (and then, from (5.6), the components of the gravitational force, $\nabla\Phi$, are independent of the space variables). (To see this, suppose ∇ is a metric connection with compatible metric g. Then with $R_{abcd} \equiv g_{ae}R^e{}_{bcd}$, one has, from the covariant constancy of X, Y and Z and the Ricci identity, $R^a{}_{bcd}k^d = 0$ for $k = X$, Y, and Z, and hence six independent solutions at each $e \in E$ to the equation $R_{abcd}F^{cd} = 0$ for the contravariant (simple) 2-form F, namely, $X \wedge Y$, $X \wedge Z$, $Y \wedge Z$, $X \wedge T$, $Y \wedge T$, and $Z \wedge T$, where T is the global vector field $\partial/\partial t$. From this, it follows that Riem vanishes everywhere on E.) Thus, the vanishing of Riem on E is equivalent, in any of the allowable coordinate systems, to the condition that the gravitational force, $\nabla\Phi$, may be simulated by the inertial force arising from a linear acceleration transformation (that is, Φ may be incorporated into the $\boldsymbol{a}(t)$ term) or, in other words, that the force $\nabla\Phi$ may be *transformed away* by a time-dependent translation. It is remarked, further, that the global vector

field $\partial/\partial t$ on E is not invariant under parallel transport around closed curves at e unless Γ is flat and that, given that Γ is not flat, no meaning can be attached to the question of whether this vector field is *orthogonal* to the space sections. It is finally remarked that the Ricci tensor associated with Riem and with components $R_{ab} \equiv R^c{}_{acb}$ can, from (5.6), be seen to satisfy [5.15]

$$R_{00} = \nabla^2 \Phi = K\rho \,, \tag{5.7}$$

with all other components zero, where ρ denotes the matter density responsible for the gravitational field and K is a constant.

5.5 Special Relativity

In this section the move will be made from classical Newtonian theory to the special theory of relativity, which was first put forward in its complete form by *Einstein* in [5.17] (although one must not forget the contributions of Poincaré and Lorentz; useful reviews of these may be found in [5.18, 19]). As a preliminary remark it may be pointed out that classical Newtonian theory (as discussed in Sect. 5.3 and in the first part of Sect. 5.4) as applied to mechanical forces and the Galilean transformations which link the inertial observers collectively satisfy the well-known classical *Newtonian principle of relativity*. However, classical Newtonian theory combined with Maxwell's electromagnetic theory does not satisfy this principle since, as pointed out at the end of Sect. 5.3, the rigid ether determines a unique *rest frame* in which the speed of light is independent of direction. If one wishes to retain a similar collection of inertial frames which are indistinguishable electromagnetically as well as mechanically, one requires different velocity addition laws between such frames and hence the Galilean transformations must be rejected. A different approach to the kinematics is thus required in which, amongst other things, this Newtonian principle of relativity is replaced by the more general *Einstein principle of relativity*, which says that inertial observers cannot be distinguished by any experiment, mechanical or electromagnetic.

This necessitates a change in the setting up of coordinates. One no longer accepts the Newtonian concept of absolute time but rather introduces a family of inertial observers, each of which has its own time coordinate. To do this one first denotes the universe of events by E as before and assumes that for such an inertial observer O a time coordinate is given as a map $T_O : E \to \mathbb{R}$. Thus, for such an observer one can still define, in an obvious way, the instantaneous space sections $S_O(t) \equiv T_O^{-1}\{t\}$ as before but which now may (and, in fact, do) depend on O. Each such section is assumed to admit a Euclidean structure and hence a Cartesian coordinate system x, y, z consistent with that structure, just as in the Newtonian case already considered. Thus O may coordinatize E by associating the coordinates x, y, z, t to an event $e \in E$ where $T_O(e) = t$. The ability to distinguish between *real* and *inertial* forces, and the concept of a free particle, as in Newtonian theory, are retained and then the spaces $S_O(t)$ are related to each other by assuming, as before, that free particles have paths given by a map of the form $t \to (x(t), y(t), z(t))$ with $x(t), y(t)$ and $z(t)$ *linear* functions of the time coordinate t obtained from the function T_O with t in some open interval of $\mathbb{R}$. It is also assumed that any such triple of linear functions is the (potential) path of some free particle. Thus each inertial observer O gives rise to a map $E \to \mathbb{R}^4$ by attaching to an event $e \in E$ the 4-tuple (x, y, z, t), where $t = T_O(e)$ and (x, y, z) are the projected Cartesian coordinates of $e \in S_O(t)$ and provide a global chart for a four-dimensional manifold structure on E. It is assumed that the collection of all charts for all inertial observers gives a smooth atlas for E. Further, one can again define particles at rest (fixed points) in O and for any fixed $t_0 \in \mathbb{R}$ one may *project* in O, as in the Newtonian case, from any $S_O(t)$ onto $S_O(t_0)$ to achieve a definition of spatial distance between any two events in E as measured in O.

In special relativity, however, two extra assumptions (in addition to the principle of relativity stated above) will be introduced. The first assumption is that if a clock is a fixed particle in O, it may be synchronized so that its time reading always coincides with the t coordinate in O of the event at which it is read, and that two clocks, one at rest in one inertial frame and one at rest in another, will, on being brought to rest in either of these frames, agree as to the *unit* (but not necessarily the actual value) of time used. Thus one has the concept of a *good clock* with clocks at rest in O being consistent with the time coordinate in O. The second assumption arises from the celebrated results of the Michelson–Morley experiment (see, for example, [5.20]). It is assumed that the

above coordinates can be chosen in such a way that if O is any inertial observer and if p and q are events in E with time coordinates t_p and t_q in O and are such that the spatial distance between them, in O, is $d(p,q)$ then the condition that a photon may pass through the events p and q is that $d(p,q) = c(t_p - t_q)$ for some constant c *and that this constant* c *is the same constant for every inertial observer* O *performing the experiment.* This second assumption, regarding the photon, shows that the constant c above is to be regarded as the speed of light in the units chosen and that this speed is the *same for all inertial observers*. Thus absolute time is abandoned and each observer sets up his own time coordinate as above. The difference between this time coordinate in O and absolute time is that, in general, if O and O' are inertial observers with time coordinate functions T_O and T'_O and e is some event, $T_O(e) \neq T'_O(e)$, and that if a clock is not at rest in O it will, in general, not agree with the coordinate function T_O on the set of events through which it passes. The exact role played by the Michelson–Morley experiment in Einstein's original development of special relativity is not entirely clear (see [5.19] for a full discussion). However, one of the main points of the assumptions of special relativity is that the ether is now abandoned as no longer being part of physics.

Continuing with the hypothesis regarding the photon, let $e \in E$ and suppose that O and O' are inertial observers such that each allocates the coordinates $(0,0,0,0)$ to e. Let $p \in E$ be any other event with coordinates (x,y,z,t) in O and (x',y',z',t') in O'. Suppose that $x^2+y^2+z^2-c^2t^2 = 0$. Then, by the photon hypothesis, a photon could pass through e and p from O's viewpoint, since the spatial distance traveled by it equals $|ct|$. Then it would also pass through e and p from O''s viewpoint and so $x'^2+y'^2+z'^2-c^2t'^2 = 0$. Now E is the manifold $\mathbb{R}^4$ so consider the tangent space T_eE to E at $e \in E$ and the bijective map $f : T_eE \to E$ obtained by mapping $v \in T_eE$ to a point of E whose coordinates are the components v^a of v in the basis $\partial/\partial x^a$ $(a = 1,2,3,4)$ for T_eE obtained from O's coordinates $x^1 = x$, $x^2 = y$ $x^3 = z$ $x^4 = ct$ (essentially the exponential map associated with the symmetric connection on E whose Christoffel symbols in this coordinate system all vanish). Then one can define a metric η on T_eE by $\eta(u,v) = \eta_{ab}u^a v^b$, where $\eta_{ab} = \mathrm{diag}(1,1,1,-1)$ and u^a are the components of u in the above basis for T_eE. One can similarly set up a metric η' at e for the observer O'. The above remarks about photons then show that the two quadratic forms η and η' share their zeros and, since they are of indefinite signature, they are proportional. One may now appeal to the principle of relativity in this theory to show that this constant of proportionality is unity. It follows that the subset $\{w \in T_eE : \eta(w,w) = 0\}$ of T_eE is independent of the observer. Denoting this subset by N_e (the *null cone* at e), one has an extra structure introduced into E at each event in E for special relativity. This structure gives rise to the *Minkowski metric tensor* and associated inner product. Thus special relativity is based on the universe E taken as the manifold $\mathbb{R}^4$ with a global metric tensor of signature $(1,1,1,-1)$ and with components $\eta_{ab} = \mathrm{diag}(1,1,1,-1)$ in the global coordinate system x^a on $\mathbb{R}^4$. The transformations which now play the role of the Galilean transformations in Newtonian theory are the Lorentz transformations; they preserve the null cone N_e at each $e \in E$ in the sense that, regarded as bijective maps $\mathbb{R}^4 \to \mathbb{R}^4$, any such map ψ is linear and satisfies $\eta(\psi(u), \psi(v)) = \eta(u,v)$ for each $u, v \in \mathbb{R}^4$. In matrix form they constitute the (Lorentz) group $\{A \in GL(4,\mathbb{R}) : A\eta A^t = \eta\}$, where the superscript t denotes a matrix transpose. Much of modern special relativity, including its nomenclature, stems from the beautiful work of Minkowski; more details, including a new translation into English of one of his major works in this area, can be found in [5.21].

One of the main changes this brings to the geometry of classical physics is the fact that in Newtonian theory (Sect. 5.4), one has (except in the trivial case when the connection Γ introduced there is flat) a metric h on the space sections but no metrical link between these space sections and the time axis. This is essentially forced upon the situation by the imposition of the absolute time concept. In special relativity, however, the absolute time concept is rejected and each inertial observer has his own time coordinate, which is assumed to satisfy a certain property with respect to photon behavior. Thus there are no longer observer-independent space sections but rather a splitting of E into space and time in a way which is observer dependent but in which the time coordinate is now metrically (in fact, orthogonally) linked to the space sections for a given observer through the Minkowski metric. The way this split differs between observers is determined by the Lorentz transformations which, essentially, map one splitting (say of O) into that of O' in such a way as to preserve the null cone. This leads to a well-defined (observer-independent) mathematical structure (the null cone) at each point $e \in E$ and hence to the metric η on E (rather than just to a metric on the space sections), which is of fundamental importance not only for special relativity but also for general relativity.

5.6 Absolute and Dynamical Variables; Covariance

Consider the situation in Newtonian theory when one is describing a situation in an *inertial* frame. The usual Newton equation relating, say, a gravitational force $\boldsymbol{F}$ on a particle of mass m to its acceleration $\boldsymbol{a}$ is $\boldsymbol{F} = m\boldsymbol{a}$. In a noninertial frame, one would have to rewrite this equation to incorporate the inertial forces that arise. However, this may be done more elegantly by writing Newton's equation in Lagrangian form and the resulting equations hold in any coordinate system. As the first step in this most elementary reformulation of the problem, and before any constraints are eliminated, the usual Cartesian coordinate system is introduced into the $3n$-dimensional configuration space (where n is the number of particles considered) and a positive definite metric is introduced on the configuration space through the kinetic energy of the system. One then proceeds in the usual fashion by the techniques of differential geometry. One achieves from the Euler–Lagrange equations an equation of motion consisting of the geodesic expression for the Levi-Civita connection of this metric augmented by the (generalized) force term. This equation reverts back to the usual Newtonian form on returning to the original inertial coordinate system where the Christoffel symbols vanish. If (holonomic) constraints are involved one may represent such constraints as a submanifold of the configuration space and rewrite the equations of motion in the (not necessarily flat) submanifold geometry. But, again, the safety of the original (Euclidean) space is still there. Standard coordinate techniques such as, for example, the relating of *ignorable* coordinates to conservation laws and the consequent reduction of the Lagrangian, introduced by Routh (see, for example, [5.22]), can be replaced by the covariant technique of seeking certain types of Killing vector fields (with respect to the above metric) on the configuration space. Everything appears covariant but one inevitably asks if anything significant is achieved by this. Of course, it has an aesthetic appeal and it is very useful in calculations. However, the covariance that seems to have been obtained is, in a sense, illusory because one can, by a definite coordinate transformation, return to the original inertial frame and the usual Newton equations. The penalty for the covariance achieved appears in the form of yet more variables (the configuration space metric) which are not dynamic in that they do not satisfy any field equations.

As another example, consider Maxwell's (source-free for simplicity) equations in an inertial frame in special relativity where the Maxwell–Minkowski tensor F, which incorporates the (observer-dependent) electric and magnetic fields, satisfies (in a standard notation; see, for example, [5.23])

$$\begin{aligned}&\frac{\partial F^{ab}}{\partial x^b} = 0\,,\\&\frac{\partial F_{bc}}{\partial x^a} + \frac{\partial F_{ca}}{\partial x^b} + \frac{\partial F_{ab}}{\partial x^c} = 0\,.\end{aligned} \tag{5.8}$$

Again, these equations hold in an inertial frame but can be made *covariant* by the simple trick of changing the partial derivative to a covariant derivative with respect to the (flat) Levi-Civita connection arising from the Minkowski metric tensor. (In fact, this change is only needed in the first equation in (5.8).) The resulting equations then hold in any coordinate system but extra variables (the metric components and Christoffel symbols in arbitrary coordinates) have been introduced into the equations. As another example, in the Cartan version of Newtonian theory (Sect. 5.4), the metric h in the space-sections but now with general components $g_{\alpha\beta}$ rather than $\mathrm{diag}(1, 1, 1)$ similarly enters the equations. In these last two examples, one can always revert back to the original metrics represented by the Minkowski metric tensor η and h, respectively, by a coordinate change and the general metric components do not satisfy any field equations having been imposed from the beginning and being uninfluenced by what is going on (that is, by the physics). Newton's *absolute* space influences physics by imposing inertial forces on anything that *dares to accelerate with respect to it*, and yet is itself uninfluenced by physics. This failure of *reciprocity* has led to these extra quantities (in these cases, metrics) being called *absolute variables* [5.12, 24, 25]. It seems that most theories can be made *covariant* by including such variables into them and which, whilst useful in many ways, does not change anything significant in the theory itself. It is in this sense that the term *covariant* has little meaning. The position vector(s) of the particle(s) in Newtonian theory which are to be determined and the tensor F in (5.8) have been called *dynamical variables* [5.24] and play a different kind of role in the theory from the absolute ones. It is these variables which satisfy field equations and in an inertial frame of reference they are the only variables which enter these equations. The Galilean transformations and those of Lorentz then preserve, respectively, the exact form of Newton's equations and the Maxwell equations (5.8) in an inertial frame, that is, they preserve the

form of the absolute variables. In the Cartan approach to Newtonian mechanics in Sect. 5.4, a Newtonian would claim to be able to identify inertial frames and thus to be able to identify a unique flat connection. In this sense, this connection is to be regarded as an absolute variable. In the Cartan interpretation the collection of frames obtained from the original one by transformations of the form (5.5) is, in this sense, *absolute*. In Newtonian–Aristotelian theory with a strict (Newtonian absolute) rest frame or Newtonian-electromagnetic theory together with a (rigid) ether (see the end of Sect. 5.3) more absolute variables are introduced, since now one has a preferred (rest frame) *observer* whose (space-time) velocity is absolute.

5.7 General Relativity

Some important points may now be drawn from the previous sections. First, it has been seen that in both Newtonian theory and in special relativity the universe of events can be regarded profitably as a four-dimensional manifold admitting special coordinate systems in which three of the coordinates take advantage of the Euclidean nature of the space sections (Cartesian coordinates) and the fourth one usefully uses a *natural* time coordinate.

Second, in discussing Newtonian gravitational theory in Sect. 5.4, it was seen how Newton's laws, together with his gravitational hypothesis and the experiments of Eötvös and Dicke, etc., showed that the path of a particle falling under the sole action of a gravitational field depended only on the initial position and velocity of the particle and not on the particle itself and that this was related to the fact that the particle's gravitational and inertial masses could be chosen equal by a simple choice of units. The equality of these mass parameters manifested itself in the fact that a gravitational field gives rise to a *well defined acceleration field* and *not* to a well-defined force field. The force field depends on the particle experiencing it at the event in question, with its common mass parameter acting as a coupling constant. Thus gravitational fields act indiscriminately on particles irrespective of their mass and make up. Now suppose that O is an inertial observer (in the sense of Sect. 5.3) and O' another observer whose motion with respect to O is, say, one of constant acceleration in a straight line. Any free particle in O's frame will now appear to have the *same* constant acceleration with respect to O', and the observer O' could interpret this as due to the existence of a certain (homogeneous) gravitational field (Sect. 5.4). This ambiguity in the description of particle motion is due to the fact that inertial forces impart the same acceleration to all particles, that is, they have the same indiscriminate action on particles as a gravitational field does. It was seen in the discussion of the Cartan connection in Sect. 5.4, that the gravitational potential was not well defined due to the fact that the inertial force arising from a translational acceleration could not be unambiguously separated from the gravitational field. Quite generally this ambiguity calls into question the concept of an inertial frame when gravitational fields are present, since it seems that one cannot separate the inertial from the gravitational forces in a satisfactory way.

Third, the questioning of the inertial frame (and observer) concept and the idea of the universe E being described by a four-dimensional manifold suggest that all observers (that is, coordinate systems) rather than those chosen for convenience (as described in the first point above) must be considered *equivalent* and that these can naturally be accommodated within some manifold structure on E. Thus, the difficulties of defining an inertial frame would lead to the concept of an inertial frame being discarded (as the ether was) because it is no longer part of physics. Similarly, the difficulty in distinguishing inertial and gravitational forces suggests that one should no longer try to do so, and thus one is led to the *Einstein principle of equivalence* of inertial and gravitational forces. This is consistent with the plan to treat all coordinate systems equally, since the (original Newtonian) differentiation between inertial frames and noninertial frames was made on the basis of the identification of inertial forces. This, in turn, suggests that the equations that determine a gravitational field, whatever they may be, should be written in such a way that they also do not discriminate between the different observers (that is, coordinate systems) and hence that they should be formally the *same* in each admissible coordinate system. Recalling the discussion of Sect. 5.6 one sees that, whereas covariance could be imposed in a somewhat *ad hoc* way if inertial frames are assumed to exist, one now has a case for covariance out of necessity. But if this covariance is achieved without the need to introduce absolute variables of the type described in that section (and which are essentially leftovers from inertial frames or other absolute objects such as the ether), it acquires a deeper meaning. Such is the principle of covariance in general relativity theory.

Fourth, and returning to the second point above, the gravitational field was seen to be able to be *transformed away* in a region if it is homogeneous. Continuing this idea (and recalling Einstein's lift experiment) the gravitational force experienced by a particle in a *general* gravitational field can be similarly transformed away *at a point* by an appropriate accelerative transformation. In some sense it may be, somewhat intuitively, regarded as a change of metric from the Minkowski metric at the point in question, but a change which will, in general, require a different coordinate transformation from point to point and not one which can be simulated by a single coordinate transformation over a region. This suggests that if a metric represents the gravitational field it may not, in general, be flat.

Fifth, one might expect that, when the gravitational field is in some sense weak, and given the success of the special theory of relativity, whatever theory of gravitation is adopted, it should *reduce to special relativity* (in some sensible way) in such circumstances. One of the main mathematical ingredients of special relativity is the null cone structure and the Minkowski metric to which it leads. Since Riemann had shown how one may describe various geometries on a manifold using metrics more general than Euclid's or Minkowski's, the suggestion is that one might incorporate the gravitational field in the geometry, that is, that the geometry itself should *reflect* or even *be* the gravitational field, relegating Minkowski's metric to the special case when the gravitational field is absent. This was the thinking behind Einstein's general theory of relativity. (It is remarkable that, at the time when theoretical physics entered a period of crisis at the beginning of the twentieth century, Riemann's work, only half a century earlier, was available in almost exactly the form that Einstein wanted it.)

Finally, given the success of Newton's gravitational theory, Einstein wanted to preserve the fact that the (differential) equations controlling the dynamical variables in general relativity theory, whatever these variables may be, should be of second order in them, as is the Poisson equation in Newton's theory. (The original work of *Einstein* can be found in [5.26, 27]).

To get some idea of how Einstein arrived at the general theory of relativity, let us return to the ideas given at the end of Sect. 5.3. Here the space slices, $S(t)$, were taken to be Euclidean. However, this choice of geometry is linked with physics only by virtue of the fact that it was assumed possible to take the preferred (straight) lines of Euclid's geometry on each $S(t)$ to be in agreement with the paths of *free* particles. This relation between geometry and physics depends, of course, on the concept and choice of a free particle. Given this, such an arrangement can be viewed as a convenient choice of straight line structure to *fit* such free particles. (It must not be forgotten that by choosing some bijective map $f : \mathbb{R}^3 \to \mathbb{R}^3$ one can, in an obvious way, map Hilbert's undefined notions of point, line, and plane, and his other concepts (incidence, betweenness, congruence, length, and angle) from one copy of $\mathbb{R}^3$ to the other, using f, to obtain a perfectly consistent geometrically isomorphic *Euclidean* geometry with a line structure which is not consistent with Galileo's law of inertia and hence not *convenient* [5.28].) However, as a serious step to some geometrization of physics it has the immediate disadvantage that Euclid's is a *homogeneous* geometry, that is, it can be represented by a manifold structure on $\mathbb{R}^4$ on which, in the sense of Riemann, a global metric is defined whose geodesics, through its Levi-Civita connection, are complete and define its *straight* lines and which is of constant (zero) curvature. Then, geometrically, any sufficiently small region of it looks the same as any other sufficiently small region. (If, for some reason, one decided to adopt the geometry of Lobachevski and Bolyai for the sections $S(t)$, similar remarks about homogeneity would still apply.) If physics and geometry interact in any way at all then, apart from the example of a rather approximate, smeared out *homogenized* physical effect of the type usually envisaged in cosmology (and which will be mentioned later) such geometries are of little value for such a geometrization. The observation that the general physical situation (usually) varies from region to region in the universe suggests that, given a wish for some interaction between physics and geometry, a more flexible form of geometry is needed than one which rests on somewhat rigid *global* axioms. (One should here, perhaps, mention the brief but penetrating remarks of *Clifford* made many years before the advent of general relativity theory [5.29].) In any case, the geometries of Sect. 5.3 are in the space sections, not on the universe E, and are thus inappropriate given Minkowski's four-dimensional work on special relativity theory and the observer-dependent nature of the time coordinate in this theory.

The type of geometry envisaged by Riemann was essentially of this more desirable form, being defined by certain functions (from the metric g and its *space* and *time* derivatives leading to its Levi-Civita connection and associated curvature) on a four-dimensional manifold M (the universe), which were sufficiently flexible to accommodate the changing physics. The local

observational physics was then described by the local metric properties of g. However, it must not be forgotten that this is a natural extension of Minkowski's four-dimensional space-time geometry and not of the three-dimensional Euclidean geometry on the space sections described in Sect. 5.4. This, together with the change in the signature of the metric from the three-dimensional Euclidean $(+,+,+)$ to the four-dimensional Lorentzian $(+,+,+,-)$ signature employed by Minkowski is not to be taken lightly as is sometimes suggested in popular works. In general relativity theory, Einstein assumed that the universe was a four-dimensional manifold of events M admitting a Lorentz metric g of signature $(+,+,+,-)$ whose Levi-Civita connection is denoted by D and the latter's associated curvature tensor denoted by Riem. This metric is an inner product on each tangent space to M, these inner products then joining together to give a *smooth metric tensor* on M. (It is *not* some *infinitesimal* distance function represented by the traditional $\mathrm{d}s^2$ although this approach has some intuitive, if indefinable, charm!) Should a change of coordinates be made in a neighborhood U of some point m of M, the components g_{ab} of g representing the metric in the original coordinates x^a change to the representative matrix g'_{ab} in the new coordinates x'^a according to the (matrix) scheme $g' = S^t g S$, where S_{ab} is the nonsingular matrix $\partial x^a / \partial x'^b$ representing the coordinate transformation. Thus, according to the well-known Sylvester law of inertia, this metric may always, by some coordinate change, be transformed to its appropriate (Sylvester canonical) form for this (Lorentz) signature, that is, the Minkowski form η_{ab} *at a point*, and, in this sense, Minkowski space is recovered (in the tangent space to M) at m. However, the Lorentz signature and four-dimensionality causes certain intuitive problems. Consider, for example, the case when the four-dimensional manifold $M \equiv \mathbb{R}^4$ is given the usual *positive definite* Euclidean metric $\bar{g}$ with global components given by the matrix diag$(1,1,1,1)$. A (topological metric) distance function $d : M \times M \to \mathbb{R}$ arises on M in a natural way from $\bar{g}$ through Pythagoras' theorem and the (natural metric) topology associated with d is the usual manifold topology on M. The usual orthogonal group of transformations $\mathbb{R}^4 \to \mathbb{R}^4$ preserves $\bar{g}$ and d in a natural way. If, however, M is now given the usual global Minkowski metric, the usual manifold topology on M cannot possibly arise from a (topological metric) distance function on M which is naturally preserved by the Lorentz group. To see this one simply considers $m \in M$, assumes that such a distance function d' exists on M, and then notes that the whole of the surface of the null cone through m must be contained in any neighborhood of m.

In some sense, the *geometry* represented by g, D, and Riem is taken as *the gravitational field* and Einstein formulated field equations for the determination of these quantities. Such equations should be expressed in such a way as to be *coordinate independent* and this was taken to mean that they should be expressed in terms of tensors on M (the principle of covariance). In this sense, there should be no preferred *observers* and so one of the main problems in Newtonian classical theory was removed. However, a little more is involved here since Einstein's equations will be seen to be differential equations for certain variables (the metric components) which are entirely dynamical and these variables will be the only variables entering the field equations. Thus, recalling the work in Sect. 5.6, no absolute variables are required and, in this sense, Einstein's theory is covariant in a much deeper way. As for the field equations themselves, Einstein apparently toiled long and hard to find them. Although no strictly deductive reasons can be given for them, some suggestions can be made which, if not compelling, are perhaps sufficient reasons for adopting them in the first instance. After this, experiment must decide and, almost a century later, general relativity is still the most successful theory of gravity available. A few guidelines can be given. Because of the success of Newtonian gravitational theory where the field variable is the gravitational potential Φ and the field equation is Poisson's second-order differential equation $\nabla^2 \Phi = K\rho$ for some constant K and matter density ρ, one supposes that Einstein's equations should also be second-order differential equations for the components of the metric tensor. In addition, the equation of motion in classical Newtonian theory of a particle of mass m and path $\boldsymbol{r}(t)$ in some *inertial* frame is then given by $m\ddot{\boldsymbol{r}} = \nabla \Phi$, where a dot denotes a (Newtonian absolute) time derivative. Again recalling the success of Newtonian theory, it may be reasonably expected that, in the event of a weak gravitational field, whichever field equations are chosen for the metric tensor should give similar results to those derived from the Newtonian viewpoint. This is justified by the remarkable success of Newtonian gravitational theory in the period between Newton and Einstein (and, in fact, after Einstein).

Recalling the discussion in Sects. 5.3 and 5.4, Newtonian theory and the principle of equivalence suggest abandoning the concept of a gravitational force in favor of well-defined gravitational acceleration at each space-time event. This defines a collection of space-time paths

for particles in a pure gravitational field, each determined by any event on it and the particle's velocity at that event. Thus, Einstein made the assumption that the equations of motion for a particle in his theory should lead them to follow (timelike) geodesics with respect to the connection D and with arc length with respect to g as an affine parameter. Similarly he assumed photons would follow null geodesics. The choice of a geodesic path (Einstein's principle of equivalence) reflects the results of the experiments of Eötvös and others, which suggest that the path of a particle in a pure gravitational field is determined by its initial position and initial velocity. In four-dimensional language this means that its path is determined by the particle's position on the space-time manifold and its (four-dimensional, timelike) *direction* in the tangent space at that point. Given that Einstein's theory is based on a metric, the geodesic is a rather natural path determined by the metric's associated Levi-Civita connection and which satisfies the conditions of the last sentence. Thus this principle of equivalence, which appears as a theorem in Newtonian theory, is now given the status of a *postulate*. In summary, the transmission of information in general relativity theory is along paths whose tangent vector is either timelike or null and are thus confined to lie on, or inside, the null cone at each point. This leads to a *causal structure* in general relativity theory, that is, an algorithm to determine how one event in space-time may influence another.

Consider now a weak gravitational field in some coordinate neighborhood U of some event $m \in M$ with coordinates x, y, z, t and, in addition, assume that it is *slowly changing* (or even constant) in time. Then one tentatively assumes that coordinates can be chosen so that the components g_{ab} of the metric g in U are very close to the values that would be obtained in an inertial frame in Minkowski space-time with metric components $\eta_{ab} = \text{diag}(1, 1, 1, -1)$, that is, $g_{ab} = \eta_{ab} + \gamma_{ab}$ where the components γ_{ab} are small compared with those of η, max $|\gamma_{ab}| \ll 1$. If one computes the geodesic equation for a particle P using the metric g and assumes that the components of its speed, computed from the rates of change of the coordinates x, y and z with respect to the proper time, are small compared with the speed of light (here unity) one obtains, by a standard calculation, the equation of motion given above in the Newtonian case but with Φ replaced (up to a multiplicative constant) by γ_{00}. Thus one might expect the field equations sought to be such that, at least under the present special restrictions, γ_{00} is constrained by an equation like the Poisson equation mentioned above for classical theory. In particular, one might ask about such a gravitational field in a region of space which itself contains no matter but which is in the vicinity of matter which is creating a gravitational field (the *vacuum field equations*). In the classical case, the Poisson equation then gives way to the Laplace equation $\nabla^2\Phi = 0$. Recalling what was said about the field equations being of second order in the metric tensor, *Eddington* [5.30] offered an argument that the only such tensor quantities must be constructed from the metric g and the Ricci tensor, Ricc, arising from Riem. Further, and regarding the possibility of a Lagrangian formulation, it is known [5.31] that the only possible second-order field equations which could arise as the Euler–Lagrange equations from some Lagrangian constructed from the metric and its partial derivatives up to any order are of the form $a\,\text{Ricc} + bg = 0$ for real numbers a and b (and a nice form for this Lagrangian can be written down). These Euler–Lagrange equations are equivalent to $cG + dg = 0$ for constants c and d, where G is the Einstein tensor, $G \equiv \text{Ricc} - \frac{1}{2}Rg$, with R the Ricci scalar. Einstein chose as his pure gravitational (vacuum) field equations the statement that the Ricci (or, equivalently, the Einstein) tensor vanishes. Thus one has Einstein's vacuum field equations

$$\text{Ricc} = 0 \quad (\Leftrightarrow G = 0)\,. \tag{5.9}$$

On substituting the above metric $g = \eta + \gamma$ into the expression for the Ricci tensor one finds that the quantity γ_{00} satisfies the Laplace equation as required in Newtonian theory for the potential function Φ. Should matter be present these equations are modified to $G = \kappa T$, where κ is a constant and T is the second-order symmetric *energy-momentum tensor* representative of the matter content of the universe. In the above approximated metric example an appropriate energy-momentum tensor could be that of a perfect fluid with zero pressure. In this case, a similar (standard) substitution into the above Einstein nonvacuum field equations reveals the Poisson equation (5.7) for Φ. The energy-momentum tensor was presumably inspired by Minkowski's writing down of the four-dimensional form of the Maxwell three-dimensional energy tensor in special relativity theory. From this Minkowski was able to unify the separate conservation laws of energy and momentum in Maxwell's theory in a single equivalent four-dimensional conservation law, which involved the divergence of this latter tensor. Since the energy-momentum tensor T above describes, in some sense, all matter sources creating the gravitational field, it should

satisfy the divergence-free (conservation law) condition $T^a{}_{b;a} = 0$, where a semi-colon denotes a covariant derivative with respect to D. Thus, from the nonvacuum field equations, the tensor G must satisfy this divergence-free property identically, which it does. It is reassuring to note that this property of G strongly restricts it as a potential choice for the left-hand side of Einstein's equations [5.31].

Some further remarks are added here regarding these field equations and the space-time (M,g). First, the metric g determines, at each point $m \in M$, the null cone of null members of the tangent space at m, and thus the possible space-times directions of photons at m is fixed. Second, since the transmission of information is taken to be along piecewise (differentiable) timelike or null paths in M then, for physics to be possible in M, one should take the manifold topology of M to be path connected (which for a manifold, is equivalent to it being connected). Third, the field equations are regarded as equations for the metric tensor g. One may then ask what actually *represents* the gravitational field, if indeed such a question makes sense. Is it the metric, the connection, or the curvature or even the sectional curvature (since it was in the latter form that Riemann first introduced curvature)? It turns out [5.28] that, in the *general situation* (and up to units of measurement), each of these quantities uniquely determines the other and so, in this sense, they are equivalent. Finally, one may ask the following question. Given that we have a space-time (M,g) are there any uniqueness theorems applying to g or D and stemming from the Einstein principle of equivalence? The answer may be given in the following way [5.32] (for an improved version see [5.33]). Suppose g and g' are smooth metrics on M with g being the original Lorentz metric on M and g' an arbitrary smooth metric on M such that the Levi-Civita connections D and D' from g and g', respectively, have the property that, for each $m \in M$ and through each member of an open subset of g- and g'-timelike directions at m, they lead to the same unparametrized, timelike geodesics on M (that is, to the same timelike geodesic particle paths in M, paying no attention to the nature of the parameters on these paths). Then if g is a vacuum metric which is not flat, it can be shown that g' is also a vacuum metric which is not flat and $D = D'$. This last equation shows that the space-times (M,g) and (M,g') agree as to what constitutes an affine parameter and hence proper time. Further, with the so-called *pp wave* metrics excepted, $g' = cg$ for $0 \neq c \in \mathbb{R}$, that is, g and g' are the same up to units of measurement. An immediate consequence of these results is as follows; it has been seen (Sect. 5.3) that the behavior of free particles in a force-free situation in Newtonian theory can be characterized by the straight lines of the Euclidean structures on each copy of S_t. In a similar way the behavior of free particles in a force-free situation in special relativity theory is characterized by timelike geodesics with respect to the Minkowski metric in Minkowski space-time. It can now be seen that for a vacuum space-time in general relativity theory which is not flat, the (unparametrized, timelike) geodesic structure, that is, the paths of freely falling free particles according to Einstein's assumption, is in this sense characteristic of the metric. This allows a *visual description* of such a situation, locally, by noting that one may, about each point of M, choose a *convex* neighborhood U which has the property that any two points of U are connected by exactly one geodesic lying in U [5.34]. Thus the local geometry in U provides a system of *straight lines* analogous to the (local) Euclidean straight lines in Newtonian theory and special relativity theory and whose resulting normal coordinates may be useful for constructing solutions of the vacuum field equations.

5.8 Cosmology

The first attempt at a *relativistic cosmology*, that is, a mathematical description of the whole of the known universe within the general theory of relativity, was made by Einstein in 1917 when the universe was believed to be (in some approximately smoothed out way) in a *static* state [5.35]. (A *Newtonian cosmology* does exist [5.36, 37] but has interpretational difficulties and, in the context discussed here, the relativistic version is the significant one.) This Einstein *static* universe laid the general foundations of (mathematical) relativistic cosmology and introduced the cosmological constant. The static solution of Einstein was shown to be physically untenable by the later discovery of the Hubble expansion of the universe but ultimately led to the discovery of more realistic cosmological solutions of the Einstein field equations and finally to the general form of the cosmological space-time metrics which are collectively associated with the names of Friedmann, Robertson, Walker, and Lemaître; the FRWL models. For a full discussion see [5.23, 36–40]. To see how these

models arose consider the extra assumptions that are introduced into general relativity theory for the purposes of cosmology. First, one needs some geometrical assumptions expressing certain symmetries that it is felt the universe possesses on the large scale and second, one must decide how to model the actual large scale physics of the universe, that is, what the smoothed out form for the energy-momentum tensor is to be. Then, finally, one imposes the Einstein field equations. Intuitively, for the geometrical (symmetry) assumptions, one takes the attitude that the vastness of the known universe (not to mention insurmountable mathematical difficulties) allows one to ignore local irregularities and to require the universe to be *homogeneous* and *isotropic*. By *homogeneous* one intuitively means that at any given *time* the universe looks essentially the same at any point in space and by *isotropic* that, at each event, there is a special *fundamental* observer for which the universe looks essentially the same in any *space direction*. However, there are obvious difficulties here because one must first decide, for homogeneity, whether any such *cosmic* time needed for its definition actually exists. In a cosmology based on Newtonian thinking [5.36, 37] absolute time would be automatically in place but there is no such equivalent in relativity theory. As for isotropy, one has the problem of saying exactly what these fundamental observers and indistinguishable space directions are. For example, it makes little sense to take these latter directions as spacelike since, according to the causality assumption associated with the null cone structure, information cannot travel to any observer in such a fashion.

Fortunately one can avoid these problems in the following way (see, for example, [5.41]). Assume that, for the isotropy condition, one means that the observational information that any observer uses in his formulation of extra (cosmological) assumptions is received in the form of photons and hence along the observer's past null cone. Then assume the existence at any point m of the space-time manifold M of an observer who cannot distinguish such directions. Now reformulate this in terms of space-time symmetries, that is, as the statement that the space-time, (M, g), in question admits a Lie algebra, $K(M)$, of *global smooth Killing vector fields* on M which is such that the isotropy algebra of $K(M)$ at m (the subalgebra of members of $K(M)$ which vanish at m and hence whose associated local transformations (local flows) fix m) is *transitive* on the null cone of null directions at m. Now each such isotropy algebra is a subalgebra of the Lie algebra L of the Lorentz group $\mathcal{L}$ and it turns out from a consideration of the subalgebras of L and of the possible orbit dimensions for $K(M)$ that this implies that this isotropy algebra is either $so(3)$ *at every point of M or L itself at every point of M*. Whichever is the case for the isotropy subalgebra at $m \in M$, such isotropy applies also to the Weyl tensor and easily shows that it must vanish at m. It follows that for each of the FRWL models (M, g) is conformally flat. These isotropies have necessary algebraic consequences for the energy-momentum tensor at m. First consider the case when the isotropy at each $m \in M$ is $o(3)$. There are infinitely many choices for this isotropy at each m and each such choice determines (and is determined by) a timelike *direction* t_m at m (the *timelike axis of rotation*). The physics, through the observations of the cosmic microwave background radiation or Hubble expansion, is assumed to determine each t_m and, once chosen, the assumption that cosmological symmetry arises through a smooth Killing *action* ensures that there exists a local, unit, timelike, smooth vector field defined on some neighborhood U of m for each $m \in M$, which spans the *axis* at each point of U. These local vector fields are then taken as the four-velocity fields of the fundamental observers. The three-dimensional subspace O_m of the tangent space T_mM to M at m and which is orthogonal to t_m at m is then the observer's *instantaneous three-space* at m. Now regard the energy-momentum tensor T as a linear map on T_mM in the usual way. It turns out that this isotropy forces O_m and t_m to be eigenspaces of T (with respect to the metric $g(m)$). Thus T takes the algebraic *perfect fluid* form and two (possibly equal) eigenvalues emerge at each $m \in M$, which together can be shown to give rise to two smooth functions on M. These functions are, from the physical interpretation of T, (combinations of) the mass-energy density ρ and pressure p of the perfect fluid represented by T and whose particles are the galaxies (or galactic clusters). Einstein's equations, with or without the cosmological constant, then provide information about the relations between the functions ρ and p. In the case when the isotropy at m is L, no such unique timelike direction t_m is determined and the whole of T_m is an eigenspace of T at m. Thus there is a unique eigenvalue of T at m and, when the physical interpretation of this eigenvalue is introduced, it is not possible to have the mass-energy density and pressure of the perfect fluid satisfying $\rho \geq 0$ and $p \geq 0$ at m unless the Riemann tensor vanishes at m. Since the isotropy would then be L at each point of M, one has, in fact, Minkowski space and a contradiction. Thus if one requires, on physical grounds, $\rho \geq 0$ and $p \geq 0$ and not Minkowski space, the isotropy must be $o(3)$ at each point of M. (If one re-

quires the isotropy to be L at some and hence any point of M, the space-time (M, g) is of constant curvature and is (locally) the de-Sitter space-time if this constant curvature is positive, the anti de-Sitter space-time if it is negative, and Minkowski space if it is zero. The Killing algebra, $K(M)$, then turns out to have dimension 10.)

Given that one takes the isotropy as $o(3)$ at one and hence each point of M, a consideration of the orbits of $K(M)$ allows one to show that either there is a single four-dimensional orbit, dim $K(M) = 7$ and (M, g) is either of constant curvature or locally of the (original) Einstein static type, or that each orbit is three-dimensional and spacelike, dim $K(M) = 6$ and, over some open dense subset of M, (M, g) is locally of constant curvature or locally of the generic FRWL type. However, there is one further point here that arises in these constant curvature situations. For these cases, and although the isotropy is $o(3)$ at each $m \in M$, extra *local* symmetries may arise over certain subsets of M, which are not accounted for in the *global* Killing algebra $K(M)$ and hence not in $o(3)$ but which have the effect of giving an isotropy isomorphic to L over this subset. This leads to the *unphysical* conditions described in the previous paragraph. If such an undesirable subset is removed the situation is much simpler and one has either a single four-dimensional orbit, dim $K(M) = 7$ and (M, g) is locally of the (original) Einstein static type, or each orbit is three-dimensional and spacelike, and dim $K(M) = 6$ and (M, g) is locally of the generic FRWL type.

For the generic FRWL case, $K(M)$ induces in the orbits a six-dimensional Lie algebra of smooth Killing vector fields with respect to the induced metric in these orbits. Thus each such orbit is itself a three-dimensional space of constant curvature in its induced positive definite metric and is thus either of the hyperbolic geometry (negative curvature), spherical geometry (positive curvature), or Euclidean (zero curvature) type. The normals to these orbits are geodesic timelike vector fields and span t_m at each $m \in M$, and their affine parameters then naturally give rise to a cosmic time function. With this cosmic time these orbits become naturally defined *instantaneous* spaces (hypersurfaces of homogeneity), and the subspaces O_m mentioned earlier are tangent to them. The fact that they are orbits of the full Killing algebra $K(M)$ ensures a well-defined cosmological homogeneity. The observers who move along the geodesics orthogonal to these hypersurfaces are the fundamental observers (mentioned earlier and originally suggested by Weyl). These observers see strict cosmological symmetry and are taken, in the physical picture, as the world lines of the galaxies (or galactic clusters). (In this sense, the fact that these spaces are orbits of the *full* Killing algebra $K(M)$ on M, rather than simply admitting a transitive Killing algebra in their (induced hypersurface) geometry from the space-time metric g is the reason for the claim that they ensure *full* cosmological homogeneity. Models with hypersurfaces admitting such symmetry with respect to their (hypersurface) geometry but which are not orbits of the full Killing algebra (and hence not cosmological in the sense meant here) can be constructed.)

For the generic FRWL space-times one may choose local coordinates t, r, θ, ϕ in which the metric takes the form

$$\begin{aligned} ds^2 = -dt^2 + R(t)^2 \big(dr^2 + &f^2(r) \\ &\times(d\theta^2 + \sin^2\theta\, d\phi^2)\big) , \end{aligned} \tag{5.10}$$

where t is cosmic time, R is a function of t only, and the function f depends on the sign of the curvature of the (constant curvature) hypersurfaces of constant t, equalling $\sin r$, $\sinh r$ or r according, as this hypersurface has positive, negative, or zero constant curvature. There is a remark to be made here. The cosmological assumption employed here rests on a *global* Lie algebra of Killing vector fields and leads to the metric (5.10) in local coordinates. One could, from perhaps a more physical (observational) viewpoint, have changed the cosmological assumption to a more local one by demanding that, for each $m \in M$ there exists a neighborhood $U \subset M$ of m and a Killing algebra $K(U)$ *on* U leading to the isotropy condition given earlier. One then arrives at (5.10) for some coordinate domain on U and one has been true to the observational physics by claiming such symmetry (represented by Killing vector fields) exists only locally. Whilst the earlier cosmological assumption certainly implies this local cosmological one, the converse is not true because the local Killing vector fields guaranteed here by this weaker assumption may not be the restrictions to each U of a global Killing algebra on M. If, however, M is simply connected, this converse is, in fact, true [5.42, 43].

The fundamental observers in cosmology see an isotropic Hubble expansion and cosmic microwave background radiation. How do these compare with the *preferred* (inertial) observers in Newtonian theory discussed in Sect. 5.3? In one sense the cosmological fundamental observers are more tightly determined than in Newton's theory since, because their velocity is fixed at any space-time point by the forced orthogonality to the hypersurfaces of homogeneity, they do not admit any

boost symmetry. But they do admit a *physical* definition since any deviation from this fixed velocity would be observable through an anisotropy in the Hubble expansion or in the temperature of the cosmic microwave background radiation. They are thus part of physics in a way that Newton's absolute time and space are not.

Current thinking in cosmology suggests that the metric of the universe satisfies local equations like (5.10) with a cosmological constant included in the field equations and with the hypersurfaces of homogeneity being Euclidean (the $f(r) = r$ case). (One thus has, at least, a local situation bearing some similarity to that in Sect. 5.3 with cosmic time identified with Newton's absolute time.) The not uncommon terminology *flat universe* to describe this model is potentially misleading! The cosmological constant is currently regarded as potentially representative of the so-called *dark energy* and the consequent *force* of repulsion between the fundamental particles. More information can be found in [5.44].

References

5.1 D. Hilbert: *The Foundations of Geometry* (Open Court, Chicago 1902)

5.2 R. Bonola: *Non-Euclidean Geometry* (Dover, New York 1955)

5.3 H. Meschkowski: *Noneuclidean Geometry* (Academic, New York and London 1964)

5.4 G.F.B. Riemann: On the hypotheses which lie at the foundation of geometry. In: *From Kant to Hilbert.* In: E. William 1996)

5.5 G.T. Kneebone: *Mathematical Logic* (Van Nostrand, London 1963)

5.6 L.M. Blumenthal: *A Modern View of Geometry* (Freeman, San Francisco 1961)

5.7 G.E. Martin: *The Foundations of Geometry and the Non-Euclidean Plane* (Intext Educational, New York 1975)

5.8 M.J. Greenburg: *Euclidean and Non-Euclidean Geometries* (Freeman, San Francisco 1973)

5.9 R. Hartshorne: *Geometry: Euclid and Beyond* (Springer, New York 2000)

5.10 H. Poincaré: *Science and Hypothesis* (Dover, New York 1952)

5.11 A. Einstein: Geometry and experience (Address to Prussian Academy of Sciences, 1921), reprinted in Sidelights on Relativity (Dover, 1983)

5.12 A. Trautman: The General Theory of Relativity, Report from the Conference on Relativity Theory, London (1965)

5.13 A. Trautman: Lectures in general relativity, Brandeis Summer Institute in Theoretical Physics (Prentice-Hall, Englewood Cliffs 1965)

5.14 E. Mach: *The Science of mechanics* (Open Court, La Salle 1960)

5.15 C.W. Misner, K.S. Thorne, J.A. Wheeler: *Gravitation* (Freeman, San Francisco 1973)

5.16 E. Cartan: *On Manifolds with an Affine Connection and The Theory of Relativity* (Bibliopolis, Napoli 1986)

5.17 A. Einstein: Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt, Ann. Physik 17, 132–145 (1905), translated in The Principle of Relativity (Dover, 1923), pp. 37–71

5.18 C.W. Kilmister: *Special Theory of Relativity* (Pergamon, Oxford New York 1970)

5.19 A. Pais: *Subtle is the Lord* (Oxford Univ. Press, Oxford 2005)

5.20 R.S. Shankland: The Michelson–Morley Experiment, Sci. Am. **211**, 107–114 (1964)

5.21 V. Petkov (Ed.): *Minkowski Spacetime: A Hundred Years Later* (Springer, Dordrecht 2010)

5.22 C. Lanczos: *The Variational principles of Mechanics* (Univ. of Toronto Press, Toronto 1966)

5.23 H. Stephani: *Relativity, 3rd edn.* (Cambridge Univ. Press, Cambridge 2004)

5.24 J.L. Anderson: Covariance, invariance and equivalence: A viewpoint, Gen. Relativ. Gravit. **2**, 161–172 (1971)

5.25 E. Kretschmann: Über den physikalischen Sinn der Relativitätspostulate, A. Einsteins neue und seine ursprüngliche Relativitätstheorie, Ann. Physik **53**, 575 (1917)

5.26 A. Einstein: Die Grundlage der allgemeinen Relativitätstheorie, Ann. Physik 49, 769–822 (1916), translated in The Principle of Relativity (Dover, 1923), pp. 111–164

5.27 A. Einstein: *The Meaning of Relativity* (Methuen, Frome London 1967)

5.28 G.S. Hall: *Symmetries and Curvature Structure in General Relativity* (World Scientific, New Jersey 2004)

5.29 W.K. Clifford: On the space theory of matter (abstract), Proc. Camb. Philos. Soc. **2**, 157 (1876)

5.30 Sir A. Eddington: *The Mathematical Theory of Relativity* (Cambridge Univ. Press, Cambridge 1965)

5.31 D. Lovelock: Mathematical Aspects of Variational Principles in the General Theory of Relativity D.Sc. Thesis (Univ. of Waterloo, Canada 1973)

5.32 G.S. Hall, D.P. Lonie: The principle of equivalence and projective structure in spacetimes, Class. Quantum Gravit. **24**, 3617 (2007)

5.33 G.S. Hall, D.P. Lonie: Projective equivalence of Einstein spaces in general relativity, Class. Quantum Gravit. **26**, 1250091–12500910 (2009)

5.34 N.J. Hicks: *Notes on Differential Geoemetry* (Van Nostrand, Princeton 1971)

5.35 A. Einstein: Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie, Sitzungsber. Preuss. Akad. Wiss. (1917), translated in *The Principle of Relativity* (Dover, 1923) pp. 177–188

5.36 H. Bondi: *Cosmology* (Cambridge Univ. Press, Cambridge 1960)

5.37 R. d'Inverno: *Introducing Einstein's Relativity* (Clarendon Press, Oxford 1992)

5.38 S.W. Hawking, G.F.R. Ellis: *The Large Scale Structure of Space-Time* (Cambridge Univ. Press, Cambridge 1973)

5.39 G.F.R. Ellis, R. Maartens, M.A.H. MacCallum: *Relativistic Cosmology* (Cambridge Univ. Press, Cambridge 2012)

5.40 J. Plebanski, A. Krasinski: *An Introduction to General Relativity and Cosmology* (Cambridge Univ. Press, Cambridge 2006)

5.41 G.S. Hall: Killing orbits and isotropy in general relativity, J. Appl. Computat. Math. 2, e130 (2013)

5.42 G.S. Hall: The global extension of local symmetries in general relativity, Class. Quantum Gravit. **6**, 157 (1989)

5.43 K. Nomizu: On local and global existence of Killing vector fields, Ann. Math. **72**, 105 (1960)

5.44 M.S. Longhair: *The Cosmic Century* (Cambridge Univ. Press, Cambridge 2006)

6. Time in Special Relativity

Dennis Dieks

We compare and contrast special relativistic time with time in prerelativistic physics. In relativity there are no unique time intervals between events, and there is no unique simultaneity relation. We discuss consequences of this situation for the philosophy of time. Although we do not subscribe to the thesis that relativistic simultaneity is purely conventional, we argue that this simultaneity is unrelated to a *flow of time*.

In this chapter we shall analyze the role and status of time in special relativity. In order to bring into relief the new features of special relativistic time, we shall first take a look at the spatiotemporal structure of classical (i. e., prerelativistic) physics. The characteristic aspects of prerelativistic time are that it is both *absolute* and *globally unique*. Classical time is absolute in the sense that time intervals between events do not depend in any way on the processes that connect these events; nor do such time intervals depend on a point of view or a frame of reference. Classical time is globally unique because whether or not any two given distant events occur at the same time is a simple matter of physical fact, according to classical physics, regardless of how far apart these events are. Classical simultaneity, thus, is a physical relation that extends over the whole of three-dimensional space and slices up four-dimensional spacetime: classical spacetime is a stack of three-dimensional spaces at-a-time. This makes it possible to speak about instantaneous states of the universe. This uniqueness of the global simultaneity relation in classical physics fits in with the intuitive notion that time *flows* via a global now that is continually shifting towards the future.

Special relativity does away with time's absoluteness and also calls into question the global character of time. The new spatiotemporal structure that replaces the classical one reflects the way time functions in relativistic physical theories: relativistic laws do not need absolute time or a unique global now, and the concept of simultaneity plays a role in relativistic physics that is much less prominent than in classical physics.

It is first of all this changed status of time that gives rise to foundational and philosophical questions in special relativity. Most importantly, relativistic simultaneity is difficult to combine with our intuitions about time flow. In the part of this chapter devoted to such issues we shall pay special attention to the relation between special relativity and the so-called *block universe* (the universe as one four-dimensional entity without a privileged now) and to the significance of special relativistic simultaneity, in particular the question of whether, or to what extent, this relativistic simultaneity relation can be said to be *conventional*.

6.1 The Spacetime of Prerelativistic Physics

In Newtonian physics space and time constitute a fixed arena in which physical processes take place [6.1]. This arena, Newtonian spacetime, is a manifold of spacetime points with definite spatiotemporal geometrical properties: any two spacetime points in it possess both a well-defined spatial and a well-defined temporal distance between them. However, this spacetime structure, introduced by Newton in his *Principia*, is richer than actually needed for classical mechanics. Since all inertial frames are equivalent for the formulation of classical mechanics, the spatial distance between events that happen at different times does not play the role of an invariant quantity in the theory (since this distance is judged differently from different inertial systems). Indeed, we can replace Newtonian spacetime by a leaner structure and still do classical mechanics by going over to *neo*-Newtonian spacetime. However, with respect to time this change does not make a difference: time remains absolute and global. The classical conception of time, characterized by these two features – which are close to everyday intuition – can serve as a foil to the special relativistic notion of time.

6.1.1 Newtonian Spacetime

When Newton formulated his mechanical laws of motion he assumed a spacetime background with a well-defined geometrical structure: Newtonian spacetime. The mechanical laws depend heavily on this definite spacetime structure. Consider, for example, the law of inertia: a body on which no forces are exerted moves uniformly in a straight line or remains at rest. For this statement to possess physical content it must be understood as to what distinguishes a straight line from a curve. Newton makes this distinction by introducing a spatial distance function between spacetime points (a straight line is the shortest connection between points, so with the help of distances curves and straight lines can be distinguished). Further, to give meaning to the notion of uniform *motion* it must not only be clear what equal distances are: a definition of equal periods of time (congruence of temporal intervals) must be available as well.

Newton posited that his spacetime is a stack of three-dimensional spaces-at-a-time, each one equipped with Euclidean spatial geometry. In addition, whether or not the temporal distance between any two pairs of instants (and, therefore, spaces at these instants) is equal is an objective feature of Newtonian spacetime. This defines a notion of temporal congruence. When spatial and temporal units have been chosen we thus have both a definite spatial and a definite temporal distance between any two spacetime points (events).

To make sense of the notion of *rest* (according to the law of inertia a body may remain at rest if no forces act on it), Newtonian spacetime must also supply a notion of *sameness of spatial position across time*. In other words, the spaces-at-a-time of which Newtonian spacetime consists must be interconnected – we can represent this by having the same spatial points, at different times, on top of each other in the stack of spaces that constitutes spacetime. In this pictorial representation Newtonian spacetime is a collection of spacetime points with a vertical *rigging*; vertical lines that connect the same spatial positions at different times (Fig. 6.1). The question of whether events at different instants occur at the same place can now be sensibly raised and answered, as can the more general question of at which spatial distance events that happen at different times occur. Because the position at which a later event occurs has a unique spatial counterpart at the instant of the earlier event, and as this counterpart has a well-defined distance to the position of the earlier event (in the Euclidean space of simultaneous events), we can define this latter simultaneous distance as being also the spatial distance between the two nonsimultaneous events we started with. With these concepts in place, a body can now be assigned an *absolute velocity* as the quotient of the traversed distance and the time interval during which this distance is covered. Absolute acceleration is similarly defined as the rate of change of absolute velocity and, finally, an inertial motion is one which does not accelerate, i. e., has constant absolute velocity.

Newton has had many critics who distrusted his notions of absolute rest and absolute motion, since these concepts did not seem to correspond to anything that is observable (interestingly, Newton's absolute time received much less criticism before the advent of relativity theory). The critics of Newton's absolute space (Leibniz, Huygens, and Mach among them) thought that it should be possible to do without such suspect notions, but they were unable to develop an alternative mechanics achieving the same successes as Newton's theory. Moreover, Newton argued that absolute space was observable after all, albeit indirectly, by means of the inertial effects that occur in motions that are accelerated with respect to absolute space – the famous bucket thought experiment furnishes a prime example of this

indirect observability. We now know that Newton was too quick with his conclusions: these experiments verify only a part of the structure of Newtonian spacetime. However, the task of giving an explanation of inertial effects in classical mechanics without invoking the complete structure of Newtonian spacetime was only accomplished rather recently, in the twentieth century, by the introduction of neo-Newtonian spacetime [6.2].

6.1.2 Neo-Newtonian Spacetime

Neo-Newtonian spacetime (also called *Galilean spacetime*) is similar to Newtonian spacetime in that the spacetime points come as a collection of instants, three-dimensional spaces-at-a-time: both Newtonian and neo-Newtonian spacetime incorporate absolute simultaneity. As has already been pointed out, this temporal ordering accords well with everyday intuitions (but is in stark contrast to what special relativity teaches us about time, as we shall see shortly). Both Newtonian and neo-Newtonian spacetime provide a temporal metric (time–distance function) that pronounces on the separation between the instants, and a spatial metric which pronounces on the separation of points *within* each instant and assigns them the structure of three-dimensional Euclidean spaces.

Newtonian and Neo-Newtonian spacetime differ, however, concerning what notions they supply for judging how the points belonging to *distinct* instants relate. Neo-Newtonian spacetime incorporates no notion of sameness of place across time and so supports no notion of absolute velocity. However, it does support a notion of absolute acceleration. Neo-Newtonian spacetime is equipped with an *affine structure*, i. e., a criterion that judges lines through the spacetime as being straight or bent, and assigns a measure of being bent. Straight lines in spacetime represent inertial motions; bent ones noninertial, absolutely accelerating motions. In neo-Newtonian spacetime, then, inertial notions are primitive, unlike in Newtonian spacetime, where they derive from the spatial and temporal metric and the preferred direction in spacetime defined by the *vertical rigging* (the notion of absolute rest) (Fig. 6.1).

The Symmetries of Neo-Newtonian Spacetime

Newtonian spacetime has spatial translations and rotations and time translations as symmetries. Neo-Newtonian spacetime, being a weaker structure, has all these symmetries plus more. Consider its depiction in the right-hand diagram of Fig. 6.1. The spacetime appears as a stack of instants which are not rigidly pinned together by a relation of absolute rest, like a loose deck of playing cards. If we displace the cards in the deck uniformly with respect to each other, straight lines piercing the deck will be mapped to new straight lines. These transformations, therefore, represent symmetries of the spacetime. Their physical interpretation is that they represent *Galilean boosts*, transformations that impart a constant velocity to everything in the universe.

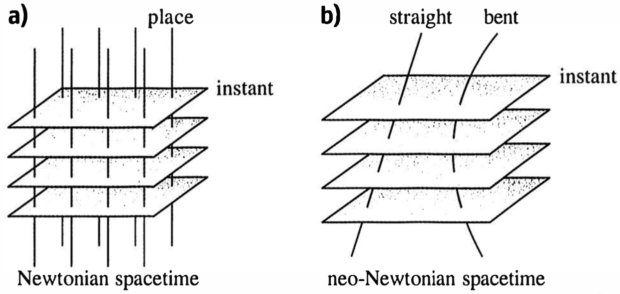


Fig. 6.1 (a) Newtonian and (b) neo-Newtonian spacetimes (one spatial dimension suppressed)

6.1.3 Classical, Absolute Time

Neo-Newtonian spacetime is the most economic spatiotemporal structure that is able to serve as a basis for classical mechanics. It does not specify a definite spatial distance between events that occur at different instants, and in this respect it is similar to Minkowski spacetime (to be discussed shortly). In order to assign a spatial distance to two nonsimultaneous events, we need to invoke something external to neo-Newtonian spacetime, namely a frame of reference (which can be represented by parallel inertial worldlines that define the state of rest in this frame of reference). Any spatial distance between two given events may result, depending on which frame of reference we choose. This is in accordance with the fact that the spatial distances between nonsimultaneous events are not invariant under the symmetry transformations of neo-Newtonian spacetime; from this it follows that these distances cannot be definable in terms of the structural properties of this spacetime alone (all these structural properties *are* invariant under the symmetry transformations). This relation between symmetry and definability (if something is definable within a certain structure, it should be invariant under the symmetries of this same structure) is helpful in judging definability questions (for example in Malament's proof about the definability of relativistic simultaneity, to be discussed later).

In contrast to the spatial distance between nonsimultaneous events, the simultaneity relation, the temporal metric and the spatial metric at-an-instant are invariant under the symmetries of neo-Newtonian spacetime. These invariant properties are the defining, *essential* features of neo-Newtonian spacetime, and therefore of time and space in classical physics. In particular, *classical time* is *absolute* both with respect to its simultaneity relation and its temporal metric. Which spacetime points can be taken to be at the same instant, and what the temporal distance is between any two events (except for the arbitrary choice of a unit) is fixed by the classical spacetime structure itself, independent of the physical processes that may be taking place in it, independent of whether or not we have introduced a frame of reference, and independent of the way in which we connect the two spacetime points. The absoluteness of these properties of time represents a core feature of classical physics.

That absolute time with these properties is essentially incorporated in classical physics can be seen from the way it occurs in the equations determining dynamical processes. Whatever path in spacetime between a given pair of spacetime points is traveled by a clock (or, more generally, by a periodic process), the equations tell us that the clock will register the time interval corresponding to the absolute temporal metric.

A *caveat* is needed here: if the clock follows an arbitrarily accelerated path, it will undergo shocks and (inertial) forces and may become distorted or damaged, maybe even to such an extent that it no longer ticks and does not register any time lapse at all. What is meant with the above statements is that absolute time is the only temporal parameter occurring in the laws of motion – a *good* clock by definition must be such that it registers this parameter. Possible distortions as a consequence of acceleration will depend on the specific design of the clock and should be corrected for (or the clock should be made more robust in order to resist the effects of acceleration). That such corrections are possible in principle is shown by the very fact that there is only one time interval that occurs in the evolution equations, regardless of the nature of the clock – this time interval corresponds to a universal feature that is independent of details about construction and materials. Therefore, all clock indications that deviate from this *true time* can be explained from the specifics of the clock, and can be corrected for. This motivates the definition of an ideal clock as a clock that measures exactly the absolute time intervals. This *caveat* is also important for our later discussion of time in special relativity: also in that case real (nonideal) clocks will generally be subject to clock-specific effects of motion, which should be corrected for (however, as we shall see, in relativity there is no unique time interval between given events even if we make such corrections).

The temporal metric of classical spacetime is thus reflected in how fast physical processes evolve according to the classical dynamics and is thus empirically accessible. Consequently, we can provide direct physical implementations of the temporal notions that are built into spacetime. For example, let us take a fiducial spacetime point and consider all possible spacetime trajectories of ideal clocks (in the sense just discussed) that have their first tick at this spacetime point. The set of spacetime points at which the clocks tick for a second time is the three-dimensional space that is one unit of time later than the starting event at which the first tick took place. All events in this three-dimensional space are simultaneous, so that we here have a direct and simple physical construction of a simultaneity hyperplane. Events simultaneous with any given event can also be constructed directly, according to classical physics, by invoking instantaneous signals (gravitational signalling).

In conclusion, classical time is characterized by a *global* simultaneity relation that slices up spacetime in a *unique* way, forming a stack of three-dimensional spaces-at-an-instant. Furthermore, the temporal distance between these instantaneous spaces is given as an objective part of the classical time structure. The temporal distance between two events is *absolute* in the sense that it does not depend on anything but the positions of the events in spacetime. In particular, it does not make a difference along which path in spacetime the time interval between the two events is measured (this will be different in special relativity). The link between the temporal properties built into classical spacetime and classical dynamics is immediate according to classical physics: time intervals and the simultaneity relation regulate physical evolution, and in turn physical processes feel and reflect the temporal structure.

6.2 The Spacetime Structure of Special Relativity

The special theory of relativity originated in the most famous of *Einstein's* 1905 *annus mirabilis* papers, *On the electrodynamics of moving bodies* [6.3–5]. In this paper Einstein set out to persuade his readers that the classical spacetime structure was not the only conceivable one, nor necessarily the one required by physics. To make his point Einstein paid close attention to how spatial and temporal intervals can actually be determined with rigid rods and clocks. He demonstrated that if his two basic postulates of special relativity are accepted, the results of such measurements necessarily deviate from what is expected in classical spacetime. Because of this emphasis on measurement, special relativity has appeared to some commentators as a theory with an outspoken operationalist flavor: a theory in which statements about space and time are reduced to statements about the behavior of rods and clocks. However, *Minkowski* showed in 1908 [6.5, 6] that the theory can also be formulated after the model of Newtonian or neo-Newtonian spacetime, namely as the specification of an independent spacetime background against which physical processes develop. But in the new, special relativistic spacetime the spatiotemporal geometric properties are different from those in classical spacetime: in Minkowski spacetime there exists a four-dimensional distance function between spacetime points, instead of one spatial distance plus one time difference as in Newtonian spacetime. In this new structure there is no longer one unique time interval between any two given events. This already suggests that the physical significance of the notion of simultaneity will be less immediate than in classical physics: there is no simple physically realizable locus of simultaneous events *one time unit later* than a given event. In fact, relativistic simultaneity becomes a *relative* notion, defined with respect to worldlines. Only in highly symmetrical configurations of worldlines (the prime example being congruences of parallel inertial worldlines, which represent inertial frames) does this lead to a global notion of simultaneity. In the case of accelerated frames of reference it is generally not possible to define a physically meaningful global simultaneity relation that is adapted to the frame.

6.2.1 Time in Einstein's 1905 Paper

At the end of the introductory section of his *On the electrodynamics of moving bodies* [6.3–5] Einstein famously declared (English translation from [6.5]):

> *The theory to be developed is based – like all electrodynamics – on the kinematics of the rigid body, since the assertions of any such theory have to do with the relationships between rigid bodies (systems of coordinates), clocks, and electromagnetic processes. Insufficient consideration of this circumstance lies at the root of the difficulties which the electrodynamics of moving bodies at present encounters.*

When Einstein subsequently starts discussing the notion of time he elaborates on the same point and warns us that a purely theoretical, mathematical description [6.5]

> *has no physical meaning unless we are quite clear as to what we understand by* time.

He goes on to explain that for the case of time at one spatial position the sought physical *definition* (Einstein's term) can simply be given as *the position of the hands of my watch* (located at the position in question). However, time thus defined is a purely local concept, and we need something more if we wish to compare times at different positions, namely a notion of distant simultaneity. In the 1905 paper Einstein briefly considers the possibility of assigning to distant events the time indicated by one fixed clock at the moment a light signal from the events reaches this clock, but immediately rejects this possibility because the time thus assigned would depend on the position of the standard clock (which would have the consequence that physical laws would become position-dependent). This consideration finally leads Einstein to the introduction of a *much more practical* procedure for synchronizing clocks. Suppose we have two clocks of the same construction, both stationary in the same inertial frame and at a certain distance from each other. Now send a light signal from clock 1 at time t_1 (as measured on clock 1), so that it reaches clock 2 at t_2 (as indicated by clock 2) after which it is immediately reflected back to arrive at clock 1 again at t_3. The two clocks can now be defined to be in synchrony if $t_2 = 1/2(t_1 + t_3)$. This is equivalent to saying that synchronicity is achieved when clocks are set such that the velocity of light, measured with their help, is the same in the direction from 1 to 2 as in the reverse direction from 2 to 1.

The frame of reference in which the clocks are stationary can in this way be provided with a global notion of time. In Einstein's words [6.5]:

> *The* time *of an event is the indication which is given simultaneously with the event by a stationary clock located at the place of the event, where this clock should be synchronous for all time determinations with a specified stationary clock.*

These and similar passages in the 1905 paper appear to be based on, and to propose, an operationalist conception of spatial and temporal notions. According to such a conception coordinates are *identified* with marks on material axes, distances *are* what is measured by rigid measuring rods, and – most importantly for our purposes – time *is* what is indicated by the hands of synchronized clocks. Indeed, Einstein's statements in these pages, and the empiricist/operationalist ideas that seem to lie behind them, have had a great influence in twentieth century philosophy of science [6.7–9]. Among the early logical positivists they constituted one of the motivations for developing the doctrine of *coordinative definitions*, according to which physical concepts (like *time*) should be coordinated, through *definitions*, to concrete physical things and procedures. In particular *Reichenbach*, in his famous book *The Philosophy of Space and Time* [6.10] (first published in German in 1928), emphasized that these coordinative definitions of physical concepts are fundamentally *conventional*. He elaborated this idea in detail in his analysis of simultaneity (about which more later). Percy Bridgman, the founder of operationalism, also took important inspiration from Einstein's 1905 paper. In his contribution to *Albert Einstein: Philosopher-Scientist*, *Bridgman* wrote [6.11]:

> *Let us examine what Einstein did in his special theory. In the first place, he recognized that the meaning of a term is to be sought in the operations employed in making application of the term. If the term is one which is applicable to concrete physical situations, as* length *or* simultaneity, *then the meaning is to be sought in the operations by which the length of concrete physical objects is determined, or in the operations by which one determines whether two concrete physical events are simultaneous or not.*

However, in his *Remarks to the Essays Appearing in This Collective Volume* [6.12], *Einstein* decidedly rejected this operationalist interpretation of special relativity and took the stance that relativistic space and time are entities in their own right, with spatiotemporal geometric properties that are independent of whether or not they are being measured. Although Einstein made these remarks more than four decades after his 1905 paper, they are probably not too far removed from the attitude that was in the background of his early work [6.8]. Indeed, to mention just one consideration that explicitly occurs in Einstein's later work but appears so physically plausible that it can hardly be assumed that Einstein thought differently in 1905: rods and clocks, and macroscopic measuring devices in general, cannot be considered as fundamental – rather, their behavior should be *explained* on the basis of microscopic fundamental laws.

However this may be, the situation was much clarified by *Minkowski's* famous 1908 lecture *Space and Time* (*Raum und Zeit*) [6.5, 6], which originated the study of special relativity as a geometrical theory of a four-dimensional spacetime manifold. According to this approach, which has now become standard, relativity theory is about an independent spacetime manifold with an in-built geometrical structure that exists even if there are no rods and clocks at all. However, before turning to this modern four-dimensional viewpoint, let us look at the characteristics of special relativistic time as they are already developed in Einstein's 1905 paper.

Einstein starts from two postulates, the *relativity postulate* and the *light postulate*. The relativity postulate asserts that all inertial systems are equivalent with respect to the form of the physical laws: in all frames of reference in which the mechanical laws hold in their standard form (without centrifugal and Coriolis forces) all other physical laws (e.g., those of electrodynamics and optics) take their standard forms as well. The light postulate says that light in empty space always propagates with the same definite velocity, independently of the state of motion of the source that has emitted the light. It is remarkable that the combination of these prima facie innocent postulates (the first says something very similar to the equivalence of inertial frames that is well-known from classical mechanics, the second is familiar from the ether theory of electromagnetism and optics) leads to conclusions that are quite staggering from a classical point of view. As far as time is concerned, the essential new conclusions are that simultaneity becomes *relative*, i. e., dependent on the frame of reference, and that moving clocks run slow with respect to stationary clocks. Both consequences follow immediately from the two postulates together with considerations about how time can actually be measured.

The relativity of simultaneity is discussed by Einstein right in the beginning of his 1905 paper, in Sect. 2 (*On the relativity of lengths and time*). Einstein asks us to imagine a rigid rod that moves uniformly, with velocity v, with respect to a given inertial system (the *stationary* system). At the two ends of the rod (A and B, respectively) there are clocks that are in synchrony with the clocks in the stationary system (i. e., they indicate the same time as the stationary clocks in whose immediate proximity they find themselves). So these clocks are synchronous from the point of view of the stationary system. Now, imagine further that an observer who is moving along with the rod performs an experiment to check whether the simultaneity condition explained above is fulfilled. That is, he sends a light signal from A to B that departs from A at time t_1 (as indicated by the clock at A). We can now calculate what time t_2 will be indicated by clock B at the instant the signal arrives there. Since clocks A and B indicate *stationary time*, the easiest way to do this is to describe the signal from the stationary frame in order to find out how much *stationary time* it uses to go from A to B. The light postulate tells us that the velocity of the light signal, as measured in the stationary frame, has its standard value c (that the light source is moving along with the moving rod plays no role by virtue of the light postulate); but of course B is moving with velocity v with respect to the stationary frame. Therefore, it takes the light a time interval $l/(c-v)$ to reach B, where l is the length of the moving rod as measured in the stationary frame. Analogously, the time interval needed by the light to go back to A is $l/(c+v)$. So we find that $t_2 = t_1 + l/(c-v)$, and $t_3 = t_1 + l/(c-v) + l/(c+v)$. Consequently, it is *not* true that $t_2 = 1/2(t_1+t_3)$; this means that the comoving observer reaches the conclusion that clocks A and B are *not* synchronized.

In other words, two clocks A and B that have been verified to be synchronous with the help of the synchronization procedure in one frame (in the above example: the stationary frame), turn out to be not in synchrony when we apply the same synchronization procedure in another frame of reference. Therefore, two events that take place simultaneously in one frame of reference (as judged with clocks that are synchronous in that frame) do not take place at the same instant according to the way time is reckoned in other frames of reference. This is the famous *relativity of simultaneity* that figures centrally in special relativity theory.

That moving clocks must run slow is also an immediate consequence of the two postulates. Suppose that we send a light signal along the y-axis in an inertial frame of reference (the *stationary frame*) and that this signal takes one unit of time (as measured by clocks in our stationary frame) to go from start (let us say in the origin of our coordinate system, $(0,0,0)$) to finish $(0,c,0)$. Now consider a second inertial system, whose origin and directions of the axes at the instant of emission of the signal coincides with those the stationary system, and moves with velocity v along the x-axis (from the viewpoint of this second frame of reference, the stationary system moves with velocity $-v$ along the axis of x). In this second frame of reference exactly the same experiment can be done: so also here a light signal is sent off from the origin in the y-direction and is made to traverse a distance c in the y-direction. By virtue of the relativity postulates, the laws governing the signals are exactly the same in the two frames (in particular, in both cases the velocity of light is c). Therefore, also the light signal in the moving frame takes exactly one unit of time, as measured by clocks that are at rest in this moving frame (comoving clocks). However, as seen from the stationary frame, the light signal in the moving frame does not go in the y-direction: its destination moves in the x-direction, so that the light must follow a slanted path to reach it (as seen from the stationary frame). Therefore, as judged from the stationary frame the light signal in the moving frame does not take one unit of time, but an interval T that follows from application of the Pythagorean theorem to triangle ABC (Fig. 6.2). We find $(Tc)^2 = (Tv)^2 + c^2$, so that $T = 1/\sqrt{1-v^2/c^2}$. So according to the time measurements in the stationary system the unit of time of the moving system is not 1 but $1/\sqrt{1-v^2/c^2}$, a number greater than 1. Clocks in the moving system must run slow by the factor $\sqrt{1-v^2/c^2}$ in order to make this consistent. Indeed, good clocks in the moving system

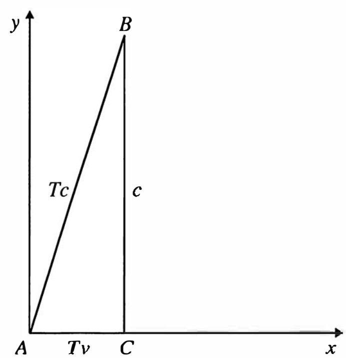


Fig. 6.2 A light signal takes longer as judged from another inertial system

all measure that it takes one time unit for the light to go up the y-axis, in the same way that good clocks of the same constitution in our *stationary* system tell us that the similar process in our system takes one unit of time (this is the relativity postulate) – this means that whatever their constitution, moving clocks must all run slower, by the same factor, than their stationary counterparts.

Of course, all inertial frames of reference are equivalent in relativity theory, so that there cannot be frames that are *objectively* stationary. Consequently, the clocks in the stationary system of the above argument must also run slow with respect to the clocks of the *moving* system: the retardation effect has to be symmetric. This is possible without contradiction because from each stationary inertial system moving clocks are compared with stationary clocks via the simultaneity relation that is defined *within* the stationary inertial system. That is, events in the life of a moving clock are compared with simultaneous events in the life of a stationary clock, using the simultaneity relation of the stationary frame. When we switch to a frame that is comoving with the moving clock (so that this frame becomes our new stationary frame), and want to compare our clock (at rest in our new frame) with the original stationary one (which is now moving with respect to us), we have to use the simultaneity relation of our new stationary frame of reference. As we have seen, simultaneity is judged differently from these different frames of reference, and this change in *simultaneity perspective* is exactly what is needed to make the retardation effect symmetric (Fig. 6.3).

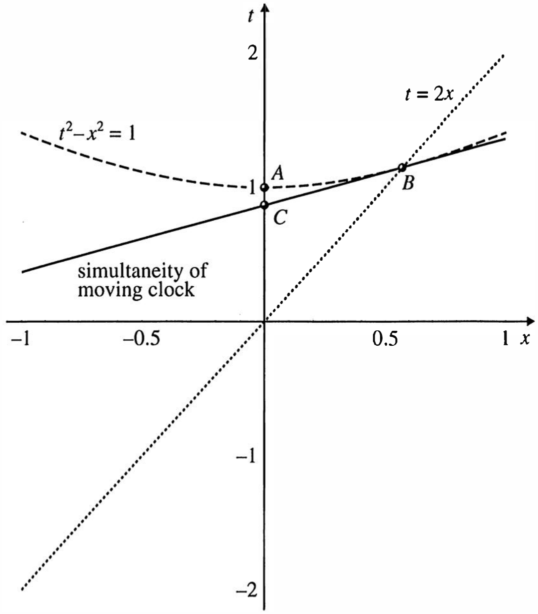


Fig. 6.3 The relativity of simultaneity makes reciprocity of the relativistic retardation consistent: a stationary clock ticks in A, which is earlier than the tick of a moving clock in B. However, the simultaneity of this moving clock makes C simultaneous with B, which is earlier than A

The slowing down of moving clocks was just explained for the case of clocks that move uniformly with respect to each other, so that they are all in inertial motion. However, the effect also applies to clocks that are in nonuniform motion. The retardation formula $\sqrt{1-v^2/c^2}$ and the above argument show that the slowing down does not depend on peculiarities of the construction of the clock, but is quite general in nature – in this sense, it may be considered a characteristic of special relativistic time itself. If a clock moves nonuniformly, small sections of its trajectory can be approximated by motion with constant velocity, and for all these (very) small parts the above retardation factor applies, when for v we take the value of the instantaneous velocity. The total retardation then follows from integration of all these local effects over the trajectory of the clock.

It is important here to remember the *caveat* of Sect. 6.1.3 about the definition of good clocks. The outlined calculation pertains to what an *ideal* clock will indicate. An ideal clock is a clock that has been corrected for discrepancies that depend on the specific construction of the clock. In this case, we should correct for effects of the acceleration. Accelerations will only affect clocks in specific construction-dependent ways – acceleration effects are not universal, in contradistinction to the relativistic retardation that only depends on v. Because of their dependence on the mechanism of the clock, on the properties of the materials that were used, and so on, the effects of acceleration can be made arbitrarily small. We can make our clocks more robust and acceleration-resistant; or we can correct their time indications for acceleration effects by means of calculations.

An immediate consequence of this slowing down of clocks also if accelerations play a role is the notorious twin effect. It is sometimes erroneously thought that the twin effect belongs to the domain of *general* relativity (since the twin case involves acceleration), but in reality it is typical of *special* relativistic time: given a pair of events P and Q, there does not exist one unique time interval between them. The amount of time that passes

between P and Q depends on the physical process that connects these two events. After having discussed the retardation effect in his 1905 paper, Einstein made the point in the following way (Sect. 4, *Physical meaning of the equations obtained with respect to moving rigid bodies and moving clocks*):

> *From this* (i. e., the retardation) *there follows the following peculiar consequence. If at the points A and B of K* (i. e., the stationary system) *there are stationary clocks which, viewed in the stationary system, are synchronous; and if the clock at A is moved with the velocity v along the line AB to B, then on its arrival at B the two clocks no longer synchronize, but the clock moved from A to B lags behind the other which has remained at B by $1/2t(v^2/c^2)$ (up to magnitudes of fourth and higher order), t being the time needed by the clock to go from A to B. It is at once clear that this result still holds if the clock moves from A to B in any polygonal line, and also when the points A and B coincide. If we assume that the result proved for a polygonal line is also valid for a continuously curved line, we arrive at this result: if one of two synchronous clocks at A is moved in a closed curve with constant velocity until it returns to A, the journey lasting t seconds, then on its arrival at A this traveling clock is $1/2tv^2/c^2$ behind the clock that stayed at rest. From this, it follows that a balance-clock at the equator must go more slowly, by a very small amount, than a precisely similar clock situated at one of the poles under otherwise identical conditions.*

It immediately follows that different paths between any two given events, traversed with different speeds, correspond to different lapses of time.

6.2.2 Minkowski Spacetime

From the relativity postulate in combination with the light postulate it follows that the description of a propagating light wave should look the same in every inertial frame. Therefore, the equation for an outgoing spherical light wave front emitted from the origin in space and time $(0, 0, 0, 0)$, $(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 = c^2(\Delta t)^2$, must hold in all inertial frames whose spatial origins coincided at the instant of the emission of the light. This mathematical expression should, therefore, be invariant in the transition from one inertial system to another. The quantities Δx, Δy and Δz stand for the distances traveled by the light in the directions x, y and z, respectively, in a time interval Δt. In effect, Einstein used this requirement of invariance in his 1905 paper to derive the equations that connect the coordinate systems of different inertial systems, the Lorentz transformations. The Lorentz transformations thus appear as the group of (linear) transformations leaving the equation $(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 - c^2(\Delta t)^2 = 0$ invariant. However, it turns out that these same Lorentz transformations have a more general property: they leave the form $(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 - c^2(\Delta t)^2$ invariant even if its numerical value does not vanish. This property of the transformation group connecting coordinate systems of different inertial frames of reference reminds one of a very similar property of the transformations in Euclidean three-dimensional geometry that connect different Cartesian coordinate systems: these transformations leave $(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2$ invariant. The interpretation of the latter invariance is of course that $(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2$ represents an objectively existing spatial distance between points, which can be represented in whatever coordinate system we wish, but whose value is independent of the choice of coordinates.

This analogy lies at the basis of *Minkowski's* proposal [6.5, 6] to interpret $c^2(\Delta ct)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2$ as a four-dimensional *distance* between points (i. e., events) in the space–time continuum. Different inertial frames of reference, with their standard spatial and temporal coordinates, thus become similar to different Cartesian coordinate systems in Euclidean spatial geometry. Moreover, just as the coordinates of different Cartesian systems sharing the same origin are linked by spatial rotations, the coordinates associated with the different inertial systems of the above wave front example are connected by Lorentz transformations representing the effects of *boosts*. (In fact, if we introduce the imaginary time coordinate $\mathrm{i}t$, with $\mathrm{i} = \sqrt{-1}$, the distance function assumes the form of a four-dimensional Euclidean distance and the Lorentz transformations appear as four-dimensional Euclidean rotations.)

In this way special relativity theory becomes a theory that places all events in the history of the universe in a four-dimensional manifold of spacetime points, Minkowski spacetime, that possesses a definite spatiotemporal geometry. This geometry derives from a distance $\mathrm{d}s$ between neighboring spacetime points: $\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2$, where x, y, z and t are inertial coordinates (this is the infinitesimal form of the distance just discussed – considering this infinitesimal form makes it possible to compute distances along arbitrary curves in spacetime). The distance $\mathrm{d}s$ is well-defined in Minkowski spacetime regardless of whether

or not material things – particles and/or fields – are present; even in empty Minkowski spacetime there thus exists a definite spatiotemporal geometry.

Like in ordinary Euclidean geometry, also in Minkowski spacetime the comparison of distances makes it possible to distinguish between curved and straight worldlines in the four-dimensional manifold. There is an important difference with the Euclidean case, though: whereas straight lines in Euclidean geometry realize the shortest distance between points lying on them, *time-like* straight worldlines in Minkowski spacetime (time-like means that $ds > 0$ along these worldlines) realize the *longest* distance between events. Straight time-like worldlines in Minkowski spacetime represent uniform inertial motion of material bodies. Straight worldlines that realize null-intervals represent rays of light (*light-like* worldlines). Curved (i. e., not straight) time-like worldlines correspond to accelerated motions of particles. Given any point in Minkowski spacetime, the light-like worldlines going through it form two cones, the future and past light cones, respectively.

Consider a pair of events that lie on a straight time-like worldline. Then there is an inertial coordinate system according to which these events occur at the same spatial position; in this coordinate system the four-dimensional distance Δs between the points is exactly $c\Delta t$. In other words, apart from a factor c, Δs is the time interval between the two events in question, as measured by a clock in inertial motion in whose life the events happen. This physical interpretation of Δs can be generalized (reminding ourselves again of the *caveat* about accelerating clocks): $1/c \int ds$ along a time-like curve connecting two events represents the time that lapses between these two events as measured by an ideal clock whose journey between the events is represented by the curve in question. The time interval that is thus defined depends on the curve along which it is calculated. This is the general expression of what we already mentioned, namely that there does not exist one unique time interval between given events, as illustrated by the twin effect. The elapsed time $1/c \int ds$ between events, which depends on the worldline connecting these events that is considered, is called the *proper time* between the events, along the connecting worldline in question.

It is an essential characteristic of the thus emerging special relativistic spacetime structure that *no global time function* is defined in it. This is quite different from the situation in Newtonian or neo-Newtonian spacetime. According to classical physics, once we have

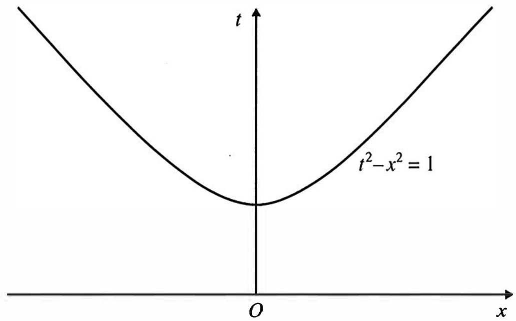


Fig. 6.4 Events one time unit later than event O fill the shaded area inside the hyperbola

chosen any particular event as our time origin, and have decided on a time unit, each event in the history of the universe can be assigned one definite time, namely the time that passes during any process that starts in the time origin and ends in the event in question. This procedure defines a time function on all spacetime points in classical spacetime. However, in relativity theory this recipe for defining global time does not work: between two events infinitely many connecting worldlines can be drawn, and the time lapse between our events depends on the worldline that is considered. The greatest lapse of time T is realized by a *straight*, inertial worldline; but any other value between 0 and T can be found by considering *curved* time-like worldlines. This, of course, is just the twin effect in a more abstract setting.

The lack of a global time in Minkowski spacetime does not lead to a problem for the formulation of physical laws. Indeed, time is still there in the form of the *duration* of processes: the time $1/c \int ds$ taken up by a process between two events occurring during its existence. It is only this time that occurs in the physical equations and is relevant for the evolution of physical systems.

This nonglobal character of relativistic time has an immediate consequence for the status of simultaneity in relativity theory. Suppose we choose an event and ask for all events that happen one unit of time later. In Newtonian or neo-Newtonian spacetime the answer is given by a three-dimensional space at-an-instant, consisting of simultaneous events all of which are one time unit later than the original event. These events cannot be mutually connected by ordinary signals with finite velocity (their propagation takes time, which would make one event later than another) but can only be linked via infinitely fast processes. In rel-

ativity theory the situation is very different, however, as we can see by considering the twin case again. By traveling fast enough (with a speed arbitrarily close to the speed of light) along a noninertial path we can push the event at which one time unit has passed arbitrarily far into the future. So the locus of events *one time unit later than a given event* does not define a sensible notion of simultaneity. Actually, these events fill the greater part of the future light cone of the originally given event. As can be easily verified, they constitute the set of events that are contained within the interior of the hyperbola with Minkowski distance 1 from the origin, $c^2(\Delta t)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2 = 1$ (Fig. 6.4). Between events in this set there is generally a time-like distance, which implies that the events in question stand in the earlier–later relation to each other and cannot be simultaneous.

So the most direct and seemingly natural physical interpretation of simultaneity fails in special relativity. In fact, more generally there is a reason to doubt the physical relevance of any relativistic notion of simultaneity: it should not be expected that simultaneity plays a significant role in determining the evolution of physical processes. Indeed, any two events that can be called simultaneous should not be connectible by a causal signal (signals cannot have infinite speeds in relativity: the maximum signal speed is the speed of light – causally connected events, therefore, stand in the *earlier–later* relation to each other). This means that in relativity theory simultaneous events are necessarily unable to have direct physical contact: any viable notion of simultaneity can only group together events that are causally cut off from each other. These events consequently cannot work together and will not function as a physically coherent whole (more precisely, they cannot do so by virtue of being simultaneous; it could be, of course, that there are relations between them because of common causes in the past). A more rigorous formulation of this same idea is provided by the observation that all physical interactions in relativistic physics are *local*. According to the relativistic equations of motion, material bodies and fields can only feel and influence each other at the spacetime points at which they are copresent. The physical changes that are brought about by interactions at a spacetime point are independent of what goes on elsewhere and only depend on the local situation. From this it follows that how one groups distant events together under the denominator *simultaneous* is immaterial for what happens in physical processes; the only thing that counts for those processes is the causal structure of spacetime. This causal structure tells us how signals propagate; propagation is local, within the future light cone of the source of the signals. On the basis of this argument one would not expect any causal role of the notion of simultaneity in the laws of relativistic physics.

Nevertheless, as we have seen, it is possible and usual to define a simultaneity relation within relativistic inertial frames of reference – Einstein's synchronization procedure does exactly this. To represent Einstein simultaneity in Minkowski spacetime, we have to realize that any inertial system can be represented by an infinite set of parallel straight time-like worldlines that correspond to *being at the same spatial position* in this inertial system. Given an inertial system and its associated notion of Einstein simultaneity, we can split up four-dimensional Minkowski spacetime as a stack of three-dimensional spaces (so that we obtain a so-called *foliation* of Minkowski spacetime); these three-dimensional spaces represent *the universe at different instants* according to the given simultaneity relation. Now, there are infinitely many possible inertial systems (each one represented by parallel inertial worldlines – a *congruence* of parallel worldlines – in a particular direction), and each of them has its own Einstein simultaneity relation. This is exactly the relativity of simultaneity that we encountered in Einstein's article. In inertial coordinates adapted to a given frame of reference the equation for space at one instant t_0 is simply $t = t_0$, so that any vector connecting two events in this space at t_0 is perpendicular to the time axis, in the following sense. Any four-vector in the space-at-a-time has the form $\boldsymbol{a} = (0, \Delta x, \Delta y, \Delta z)$, and any four-vector along the time axis has the form $\boldsymbol{b} = (\Delta t, 0, 0, 0)$, so that $\boldsymbol{a.b} = 0$ if we define the *Minkowski inner product* $\boldsymbol{a.b}$ by $\boldsymbol{a.b} = ca_1b_1 - a_2b_2 - a_3b_3 - a_4b_4$. Since the value of this Minkowski inner product is invariant under Lorentz transformations, we find that in arbitrary coordinates $\boldsymbol{a.b} = 0$ if $\boldsymbol{a}$ points in the direction of the simultaneity hyperplane and $\boldsymbol{b}$ points in the direction of the congruence of inertial worldlines that represents the frame of reference. In other words, Einstein simultaneity in a frame of reference corresponds to Minkowski orthogonality with respect to the inertial worldlines characterizing *rest* in this frame of reference.

Since Einstein simultaneity is relative (it corresponds to Minkowski orthogonality *with respect to chosen worldlines*), it is obviously nonunique. Minkowski spacetime can be foliated in infinitely many ways using simultaneity relations adapted to different frames of reference. This is completely different from the situation in prerelativistic spacetimes: both Newtonian and neo-

Newtonian spacetime possess an in-built and unique foliation as a stack of spaces-at-an-instant, provided by absolute simultaneity. As we have seen, this foliation is physically significant according to classical physics: a clock traveling at arbitrary velocity and acceleration will register the number of simultaneity hyperplanes its worldline crosses: the next tick will occur at the event at which its world line pierces the simultaneity hyperplane that lies one time unit to the future. There is no corresponding general physical significance, for clocks in arbitrary states of motion, of the simultaneity hyperplanes in special relativity.

6.2.3 Simultaneity in Noninertial Frames of Reference

We have emphasized two key features of time in special relativity: first, time intervals between events are no longer unique but have become relative to the processes connecting the events; and second, the simultaneity relation introduced by Einstein is relative to *inertial frames*. One may wonder what the special role of inertial frames in this context is. Would it be possible to generalize the considerations about simultaneity to noninertial frames? If not, what distinguishes inertial frames from other frames with respect to the role of time?

To investigate this question it is useful to look at the situation in a rotating frame of reference [6.13]. In terms of polar coordinates attached to a frame that rotates in the $x-y$ plane of an inertial system, with angular velocity ω, the Minkowski distance $\mathrm{d}s$ assumes the form (we suppress the z coordinate)

$$\begin{aligned}\mathrm{d}s^2 &= (c^2 - r^2\omega^2)\,\mathrm{d}t^2 - \mathrm{d}r^2 - r^2\,\mathrm{d}\varphi^2\\ &\quad - 2\omega r^2\,\mathrm{d}\varphi\,\mathrm{d}t\,,\end{aligned} \tag{6.1}$$

with t the time coordinate of the inertial system. As we know, $\mathrm{d}s/c$ represents the time measured by an ideal clock whose worldline connects the events between which $\mathrm{d}s$ is calculated. This entails that a clock at rest in the rotating frame will indicate the time

$$\frac{\mathrm{d}s}{c} = \sqrt{\left(1 - \frac{r^2\omega^2}{c^2}\right)}\,\mathrm{d}t\,. \tag{6.2}$$

Equation (6.2) implies that clocks at rest in the rotating frame run slow compared to clocks in the inertial frame, as expected ($r\omega$ is the velocity of the clock with respect to the inertial frame).

When we now turn to simultaneity we find, also as expected, that $\mathrm{d}t = 0$ does not correspond to (local) Einstein synchrony as judged from the rotating frame. Remember that the definition of Einstein synchrony of two clocks A and B is that a light signal sent from A to B and immediately reflected to A, reaches B when B indicates a time that is halfway between the instants of emission and reception, respectively, as measured by A. Suppose that A and B, both at rest in the rotating frame, have positions with coordinate differences $\mathrm{d}r$ and $\mathrm{d}\varphi$. A light signal between A and B follows a null worldline

$$(c^2 - r^2\omega^2)\,\mathrm{d}t^2 - \mathrm{d}r^2 - r^2\,\mathrm{d}\varphi^2 - 2\omega r^2\,\mathrm{d}\varphi\,\mathrm{d}t = 0\,. \tag{6.3}$$

This equation gives the following solutions for $\mathrm{d}t$ when it is applied to the signals from A to B and back, respectively,

$$\mathrm{d}t_{1,2} = \frac{\pm\omega r^2\,\mathrm{d}\varphi + \sqrt{(c^2 - \omega^2 r^2)\,\mathrm{d}r^2 + c^2 r^2\,\mathrm{d}\varphi^2}}{c^2 - \omega^2 r^2}\,. \tag{6.4}$$

If t_0 is the time coordinate of the emission event at A, the event at A with time coordinate $t_0 + 1/2(\mathrm{d}t_1 + \mathrm{d}t_2)$ is Einstein simultaneous with the event at B with time coordinate $t_0 + \mathrm{d}t_1$. It follows that Einstein synchrony between infinitesimally close events corresponds to the following difference in the t-coordinate

$$\mathrm{d}t = \frac{\omega r^2\,\mathrm{d}\varphi}{c^2 - \omega^2 r^2}\,. \tag{6.5}$$

Therefore, for events that differ in their φ-coordinates $\mathrm{d}t = 0$ is not equivalent to Einstein simultaneity. This was to be expected because of the motion in the tangential direction with respect to the inertial system.

The important thing is that expression (6.5) demonstrates that Einstein simultaneity between neighboring events in the rotating frame cannot be extended to a *global* notion. Indeed, if we go along a circle with radius r, in the positive ϕ-direction, while establishing Einstein simultaneity along the way, we create a *time gap* $\Delta t = 2\pi\omega r^2/(c^2 - \omega^2 r^2)$ upon completion of the circle. Doing the same thing in the opposite direction results in a time gap of the same absolute value but with opposite sign. So the total time difference generated by synchronizing over a complete circle in one direction, and comparing the result with doing the same thing in the other direction is $\Delta t = 4\pi\omega r^2/(c^2 - \omega^2 r^2)$.

Now suppose that two light signals are emitted from a source fixed in the rotating frame and that they start traveling in opposite directions along the same circle of constant r. Suppose further that we follow the two signals while locally using standard synchrony; this implies that locally the standard constant velocity c can be attributed to the signals. We therefore conclude that the two signals use the same amount of time in order to complete their circles and return to their source, as calculated by integrating the elapsed time intervals measured in the successive local comoving inertial frames (the signals cover the same distances, with the same velocity of light, as judged from these frames). However, because of the just-mentioned time gaps the two signals do not complete their circles simultaneously, in one event. There is a time difference $\Delta t = 4\pi\omega r^2/(c^2 - \omega^2 r^2)$ between their arrival times, as measured in the coordinate t. This is the celebrated Sagnac effect.

The (experimentally confirmed) occurrence of this Sagnac effect is the empirical counterpart of what we just derived, namely that the locally defined Einstein simultaneities on the rotating disc do not mesh; they cannot be combined into one global Einstein simultaneity that is everywhere adapted to the rotating system. This purely local significance of Einstein simultaneity is a quite general feature of accelerated frames of reference (although there are exceptions, like the case of hyperbolic motion – but even here Einstein simultaneity cannot be made to extend over the whole of space). Inertial frames stand out by their symmetry: they are characterized by *parallel* straight time-like worldlines, so that the spatiotemporal homogeneity and symmetry of Minkowski spacetime is respected in them. This is responsible for the fact that no Sagnac time gap exists in inertial frames. The *symmetry* (homogeneity and isotropy) of inertial frames makes Einstein simultaneity into a global notion in these frames. By the same token, the resulting global simultaneity hyperplanes supply symmetrical ways of foliating Minkowski spacetime, but do not possess an obvious significance in terms of a dynamically progressing now, causality, or a flow of time.

6.3 Philosophical Issues

Minkowski's four-dimensional world is a representation of the whole of history *at once*, in one picture. This *block* seems to do away with traditional conceptions about temporal becoming. In the four-dimensional block universe all events, past, present and future, are where they are and cannot change; they are fixed, and this may suggest a form of determinism.

In this section we shall address these and similar philosophical issues relating to relativity theory. Another important group of questions concerns the status of relativistic simultaneity. As we have seen, in Minkowski spacetime simultaneity becomes a relative notion since it is defined relative to worldlines: there is consequently no *unique* global simultaneity relation in Minkowski spacetime. Each inertial system possesses its own Einstein simultaneity. If we attempt to define Einstein simultaneity with respect to a collection of noninertial worldlines it is even generally impossible to arrive at a global relation at all. Moreover, there is a further issue: even in a given inertial frame of reference one may doubt the uniqueness of simultaneity. This is because simultaneity was introduced via a *definition* that *stipulated* a synchronization procedure (think back of Einstein's proposal for synchronizing clocks) – and one definition or stipulation may be replaced by another. This line of thought leads to the notorious *conventionality of simultaneity thesis*, which has been debated ever since Einstein's 1905 paper.

6.3.1 Relativity and the Block Universe

In Minkowski's version of special relativity center stage is taken by the four-dimensional spacetime manifold and its geometry. In so-called Minkowski diagrams events are represented as points, in the same way as spatial points in three-dimensional analytical geometry – the only difference being that now a time axis has been added to the three spatial axes. In a complete Minkowski diagram of the universe every event in the universe's history is included, each at its own spatial position and time of occurrence. Of course we cannot actually draw and specify this diagram representing all of history, since we know only a very tiny fraction of it. However, this epistemological consideration does not provide a valid argument against the *existence* of the complete four-dimensional picture. Indeed, the diagram exists by virtue of the fact that past, present, and future exist in the harmless sense that all past events took place at their places and times in history where they actually took place, present events are now taking place at their

individual positions, and future events (whatever they will be) will take place at their own positions and instants. The history of the universe *is* this collection of all events in past, present, and future, and the four-dimensional Minkowski picture is defined as their one-to-one representation. Assuming that this picture exists is equivalent to assuming that the universe has a unique history. So the only supposition made in asserting the existence of the complete Minkowski diagram is that the universe possesses an actual and well-defined history. In what follows we shall accept this premise.

The four-dimensional picture is one entity, it is there *at once*: we could imagine it as being before our eyes at one particular time, as one *block*, the *block universe*.

This four-dimensional block is usually introduced in the context of relativity theory, like we just did. However, it should be noted that the possibility of four-dimensional representations is independent of the validity of relativity theory, and that these or similar representations are, in fact, widely used also outside of physics. Any history book specifies events at different places and times – such a historical account is presented to us as one whole. It is wholly present to us at one instant, and its truth, in the sense of its one-to-one correspondence with what actually has happened, is independent of what time it is now. The same applies to television guides or railway timetables. Obviously, the possibility of these and similar examples is independent of the validity of relativity theory and the appropriateness of Minkowski spacetime; it only depends on the localizability of events by means of three spatial coordinates and one time coordinate (in any given global coordinate system). In classical Newtonian physics (parts of) the history of the universe can also be represented in a four-dimensional picture, in this case by means of events placed in Newtonian spacetime. Therefore, block universe representations are not confined to relativity theory or even to physics in general. Still, relativity theory, with its nonuniqueness of time, adds new and significant elements and is often adduced as strong support for the idea that the entire history of the world *should* be seen as one thing, wholly present at once. That is, relativity theory is often put forward as providing physical support for the philosophical doctrine of *eternalism* [6.2, 14–17].

Eternalism is the position, in the philosophy of time, according to which all events in history are equally real: they all *exist*, at their own positions and times, without there being an absolute distinction between past, present, and future. Any such distinction should according to the eternalist be interpreted as relative and indexical: what is past, present, and future depends on the position in spacetime at which the statements involving these temporal distinctions are made. So for Aristotle we are in the future, whereas Aristotle is in the past for us; but according to eternalism there is no *absolute* past, nor an *absolute* future. Accordingly, it does not make sense to say that Aristotle is in the past *tout court*, without the specification of a point of reference.

The position opposite to eternalism is that of *presentism* [6.2, 14, 16, 18], which says that there *are* objective distinctions between the modes of being of events in the history of the universe: the present – the now – is *real* in a way that is objectively different from the way the past and the future exist. Usually presentists argue that the future does not (yet) exist at all, and that the past has ceased to exist, so that reality is confined to the present.

Besides eternalism, another conclusion that is sometimes drawn from the use of the four-dimensional representation is that it implies the absence of change: the block is *static*, one changeless entity. The entire history of the universe is included in it, and there is no possibility of making it different from what it is.

This fixedness has also given rise to the notion that the four-dimensional block picture leads to *determinism* (see [6.15] for a discussion and [6.19] for a defense of this idea). The background of using the term *determinism* is that events are completely determined by how they are included in the block; e.g., it is impossible to change the future, since it is already indicated in the Minkowski diagram what it will be. All events in history are completely fixed in this same way, and this can only be so if there is determinism (according to those who defend this idea).

It should be noted, however, that ambiguities have crept in, both in the use of *determinism* and of *static*. Determinism in the sense in which the term is used in physics is about whether the equations of motion have unique solutions once boundary and initial conditions have been specified. In a deterministic universe, according to this definition, data on a *Cauchy hypersurface* (a set of points from which predictions can be made, like three-dimensional space at one instant) completely fix events elsewhere in the universe via the laws of motion. Therefore, *physical determinism* is a doctrine about the *relation*, specified by the physical evolution equations, between what happens at different times. However, in the *block determinism* that was just explained there was no mention at all of evolution equations or even of physical theory. This block determinism is consequently completely different from physical de-

terminism. It is not about physical relations between events, like physical determinism is; but it is about the relation between events in the history of the universe on one hand, and their representation in the block universe on the other. If the four-dimensional Minkowski picture of the world is accurate and faithful, history cannot be different from what the representation says it is. The *cannot* here expresses *logical* necessity; it is the *definition* of faithfulness of a representation that is responsible for the fact that there cannot be discrepancies between a faithful representation and what is represented. There is no connection at all here with physical determinism or causality. The future, and the past, are fixed and determined in the block determinism sense because they cannot be different from what they will actually be (in the case of the future) or from what they actually were (in the case of the past). This is tautological. In accordance with this diagnosis, it makes no difference for representability in the form of a four-dimensional block whether the history of the world is governed by deterministic equations or is subject to stochastic evolution. In both cases, there will be exactly one future, although this future is not fixed by present conditions and physical laws in the case of stochastic evolution. This is immaterial in the present context; the existence of one actual future is enough to *determine* the block.

In the case of the use of *static* as a characterization of the block universe more confusion is lurking. The block universe as an entity in itself evidently cannot change; this, again, is a logical necessity. Indeed, the history of the universe cannot be different from what it is, nor can its faithful representation be different from what a faithful representation of the actual history is, so the four-dimensional block comprising the total history cannot change in any way. But this does not exclude the existence of change *within* the universe. The events represented *in* the block are, generally speaking, part of dynamic processes, e.g., the motion of objects. The change that is inherent in such processes is fully included in the block, in the case of an object's motion by the direction of its worldline, and more generally by the attribution of different properties at different instants. So although it is true that the block per se is changeless, by definition, this implies nothing about the presence or absence of physical change in the universe. So it seems a non sequitur to conclude from the existence of the block representation that there is no change in the world. However, we have not yet gone to the heart of the matter and should pay more attention to the role of relativity theory in this debate.

As we have already noted, four-dimensional representations are possible even in the case of Newtonian physics. However, in that context comments about a lack of change or the future being fixed are never heard. This is because in the prerelativistic case the block universe as one whole can be considered a kind of summarizing overview of history that leaves out important structural details. To have a full view of history, according to classical physics, one should include in the picture that there is an absolute, unique, and global simultaneity relation that slices up the block. The classical block can, therefore, be thought of as a continuous stack of spaces-at-an-instant, and this makes it possible to combine this block with traditional and intuitively appealing notions about time. In particular, we may think of the classical block as *growing* by adding layers of new history [6.14, 16, 20], if we subscribe to the so-called A theory of time.

According to this A theory time is dynamic, in the sense that it *flows*, with a present that *moves* from past to future [6.21]. This is to be contrasted with the B theory of time according to which a complete description of the temporal evolution of the world can be given while using only relational terms like *earlier than* and *later than*. The A theory is closer to our direct experience of time and our intuitions, but it faces severe difficulties in giving clear meaning to the concept of *motion* of the now (ordinarily, when we speak of flow or motion we mean change of position *in time*; what, then could motion of time itself mean?).

It would be beyond the scope of this chapter to discuss the pros and cons of the A and B theories of time – for our purposes here it suffices to notice that the A theory fits the classical block, corresponding to Newtonian or neo-Newtonian spacetime, much better than the relativistic block. This is because Minkowski spacetime does not possess a unique global simultaneity structure. So even if clear sense could be made of the notion that the now progresses, application of the A theory to relativistic spacetime still faces the difficulty of deciding on which of the infinitely many Einstein simultaneity hyperplanes the now has to be located. Special relativity denies that one of these simultaneity hyperplanes can be considered as privileged; this is exactly what is said by the relativity postulate (according to which all inertial systems are equivalent).

An alternate way of showing the difficulty of combining the A theory of time and presentism with relativistic simultaneity makes use of the observation that relativistic simultaneity is *not transitive*, in the following sense. If event A is simultaneous with event B

according to the simultaneity of frame K, and B is simultaneous with C according to the simultaneity of K', A will generally not be simultaneous with C, neither according to K nor according to K'. This is no problem, and even natural, within special relativity itself. However, if we want to say that simultaneity hyperplanes represent what is present or what is real, in an absolute sense (i. e., without relativizing to a frame of reference) we obviously run into trouble [6.19, 22]. These absolute notions of presentness and reality are meant to be transitive: e.g., if B is equally real as A, and also equally real as C, then A and C must also be equally real. The interpretation of relativistic simultaneity in terms of an absolute present that represents reality, therefore, leads to contradictions.

So there is a fundamental tension between relativity theory and traditional A conceptions of time. This is the real background of the intuitive complaint that the block universe is *static*. Philosophers who subscribe to this complaint do not say that there literally is no motion or change in the universe; they admit, for example, that a particle can be at different positions at different instants of time. However, they do maintain that there is no *real change* in the block, in the sense of an objectively moving dynamic now that separates past from future.

One natural way of responding to this is to accept a B theory of time and to deny that there exists *real change* in the sense just explained. We then can accept the block universe, with all the temporal relations between events built into it, as a complete description of the universe's history. This implies eternalism and rejection of presentism. We have to be careful, however, about how we characterize this eternalist position. It would be very misleading to say, e.g., that according to eternalism the future *already exists*. The future, as judged from our position in spacetime, is not now but at later times, also according to the eternalist. So the use of *already* in its ordinary temporal sense is inappropriate. What can be said, however, is that in eternalism there is no *absolute* distinction in terms of past, present, and future between events. According to the eternalist all events occur at their own places and times, and there are temporal relations of *earlier* and *later* between them, but the block universe does not contain any structure on the basis of which it would make sense to say *the universe is now at this or that stage of its evolution*. In other words, all references to past, present, and future become indexical: they need a point of reference relative to which they can be evaluated and become definite. A task for the block view, with this B theory of time, is to explain our immediate experience of time including our feeling that the present is continually slipping away [6.23].

Another way of dealing with the lack of a unique global present is to add to the block a set of privileged simultaneity hyperplanes by hand, perhaps motivated by the expectation that this will be justified by a more general physical theory (like general relativity or a future successor of it). It may then be hoped that these added simultaneity hyperplanes can play the role of nows in an A theory of time [6.24–26]. Within the domain of special relativity itself this manoeuver is clearly ad hoc, however. In the wider context of general relativity the situation is controversial: the general relativistic field equations do not contain any reference to a notion of global simultaneity, but it is true that certain cosmological *solutions* of the field equations, with symmetrical distributions of matter and energy, do allow for a natural foliation and an associated notion of global simultaneity. It remains unclear, however, whether these foliations singled out by material symmetries can be said to have any conceptual connection with the *flow of time* [6.27] (compare our later discussion about the relation between simultaneity and symmetry).

Finally, one may attempt to adapt A notions of time to make them compatible with special relativity. One way of doing this is to *relativize* the notion of the flow of time and the idea of presentism, by making the now *frame dependent*. In its application to presentism this approach leads to the consequence that what is *real* (namely the present) itself becomes frame dependent [6.2]. This would circumvent the transitivity argument that we encountered above. However, a *relativity of reality* is so far removed from the intuitive background of presentism that this manoeuvre does not seem very attractive to many presentists.

This relativizing of A concepts keeps in place the idea that the now is spatially extended, and that the universe can be seen as a continuous succession of global nows. A more radical idea for modifying A concepts is to do without distant simultaneity at all, and thus to break loose from the problematic characteristics of this concept in relativity theory. This can be done by thinking of the present not as spatially extended, but as point-like; in this case each event defines its own present, and does not share this present with any other point-presents [6.27]. The flow of time may then be construed as taking place along causal processes, like the worldlines of particles [6.28]. This localized notion of time flow seems the most promising if we wish to combine A type notions of time with special relativity. However,

it does not unambiguously define a global present and is rather distant from the original A intuitions, just like the previous proposal. Moreover, the problem of giving meaning to the *motion* of the present remains undiminished in this version of the A theory (as in all other versions of it).

6.3.2 The Conventionality of Simultaneity

In Sect. 1 of his 1905 paper Einstein discusses the *definition of time*. He describes a situation in which we have two clocks of identical construction, both at rest in the same inertial system, the first at position A and the second at position B. Einstein comments that observers at A or B, or observers located in the immediate vicinity of these points, would have no problem in assigning a time to events occurring in their neighborhoods; they can simply tell the time by looking at their nearby clocks. However, Einstein continues, this way [6.5]:

> *We have only defined an* A-time *and a* B-time, *but no time that is common to A and B. This latter time can now be defined if we stipulate* by definition *that the time needed by light to go from A to B is equal to the time needed to go from B to A.*

The emphasis here of *by definition* is Einstein's own, in the original paper. Einstein clearly wants to draw attention to the fact that the temporal notions that he has introduced thus far (the A-time and the B-time) do not suffice to *compare* times at different locations – for this temporal coordination we need a relation of *simultaneity* that relates the instants of events taking place at a distance from each other. Of course, it is exactly this simultaneity relation that will turn out, a little later in Einstein's article, to be necessarily different from the classical one – this result holds the key to relativity theory.

If we only have local A and B-times we cannot determine the speed of any signal, because we are not able to compare its time of departure with its time of arrival. If we did know the speed of some signal, for example that of light, the problem would disappear because we could simply synchronize clocks by sending a light signal from one clock to another and by taking into account that this signal takes a time L/c to reach its destination (with L the distance between the clocks and c the speed of light). However, given that we cannot yet determine signal velocities without being able to ascertain simultaneity, we end up in a circular argument unless we brake the impasse by deciding on a concrete procedure for synchronizing clocks. We need a *definition* of how to proceed, and this definition will determine both simultaneity and the speed of light and other signals. This is what is achieved by Einstein's stipulation about the equal velocities to and fro between A and B. Note that the *round trip velocity* of the light, between A and B and back again, can already be measured with *one* clock, without simultaneity. Adding Einstein's rule now fixes the *one-way* velocities, namely as being equal to the round trip velocity. This, in turn, completes the definition of simultaneity, for we can set the clock at B at $t_1 + L/c$ when the light arrives at B, given that it departed from A at time t_1 (as indicated by the clock at A) and given that the one-way velocity between A and B is c.

The just-discussed passage in Einstein's 1905 paper has given rise to a notorious philosophical debate [6.29, 30]: is simultaneity in special relativity *factual* or rather *conventional*? The immediate reason for this debate is Einstein's use of the expression *stipulate by definition* (*festsetzen durch Definition*). Stipulations and definitions cannot be true or false, but rather are the results of our decisions; they are conventions. Einstein's procedure, therefore, appears to determine both simultaneity and the value of the speed of light on the basis of a convention and not on the basis of physical facts. Even after we have made our (conventional) choices for units of time and length it is according to the conventionality thesis a matter of our decision, and not something already decided by nature, what the speed of light in any given direction is. Only the round-trip velocity has an objective status, since it can be determined without invoking simultaneity.

Reichenbach has given a systematic and influential further elaboration and explanation of this conventionalist position [6.9, 10]. Reichenbach's core argument is epistemological in character: we do not have immediate empirical access to distant simultaneity. Because of the distance between the events that we wish to compare, and the fact that we can only make *local* observations in a direct way, we need a stipulation of the sort that Einstein made in addition to our direct observations. According to Reichenbach, local clock indications *are* factual, since they consist in spatial coincidences of material objects, like the coincidence of a pointer with a mark on a dial, and these are things that are open to immediate observation if we find ourselves at the positions in question. These observations do not depend on conventions (except in the trivial sense that we think up words to describe them, choose units in order to number the marks on the dial, and decide to focus our attention on these things in the first place). Whether or

not the hands of a clock touch a certain mark on the clock's dial is not stipulated by us but is given to us by nature. In contrast, simultaneity cannot be perceived: it has to do with a comparison between distant locations and we cannot be at two positions at once. We accordingly need some rule to tell us how to establish simultaneity on the basis of facts that *are* observable, and it is this rule that gives empirical content to statements about simultaneity.

A refinement of this theme, immediately given by *Reichenbach* in his discussion of relativistic simultaneity [6.10], is the so-called *causal theory of time.* Here it is added that it is factual that along a causal chain time progresses, with the consequence that the time of arrival of a signal is objectively *later* than its departure. This addition is in the same epistemological spirit we just explained: in principle we could travel along with the signal and observe directly that time passes.

The causal theory of time makes earlier–later relations between events that are causally connected, or could be causally connected, factual. However, this earlier–later structure does not determine a unique simultaneity relation. From this Reichenbach concludes that simultaneity is not factual but *conventional.* Accordingly, it is up to us to choose a simultaneity definition, and this can be done in many different ways. These choices give rise to descriptions that look different, but actually are equivalent: they all possess exactly the same empirical content.

Consider the situation with one clock at A and one clock at B again. Suppose we send a light signal from A to B at A-time t_1, and after the signal has arrived at B, at B-time t_2 say, we reflect it immediately to A where it is subsequently received at A-time t_3. Einstein's simultaneity definition is equivalent to saying that the clocks at A and B are in synchrony if $t_2 = 1/2(t_1 + t_3)$. However, says Reichenbach, we also could use the synchronization rule that stipulates that the clocks are synchronous if $t_2 = t_1 + \epsilon(t_3 - t_1)$, with ϵ any real number between 0 and 1. According to this alternative rule the to and fro velocities of light between A and B are no longer equal; they are $c/2\epsilon$ and $c/2(1-\epsilon)$, respectively. Simultaneity is now clearly judged differently than in Einstein's proposal, and the one-way velocities of light have also become different, but the description is empirically equivalent to Einstein's because the round trip velocity of the light, measured with one clock, is still $2L/c$. This is the only quantity that is directly accessible to observation.

Some of these empirically equivalent descriptions are simpler than others. However, according to Reichenbach this *descriptive simplicity* only yields a *pragmatic* argument for preferring one definition of simultaneity over another – pragmatic arguments relate to our interests and preferences, but not to truth. Thus, Reichenbach admits that the definition of simultaneity that makes the speed of light the same in all directions is simpler for us, easier to remember, and more readily applicable than alternatives. However, this does not mean that alternatives have a lesser claim to being true. As long as they leave the local facts and later-than relations along causal chains the same, all these theoretical schemes are equally true or false.

Reichenbach does note, however, that it is an objective fact – independent of our decisions – that a choice of simultaneity that leads to equal one-way speeds of light, and isotropy in general, *is possible.* He maintains that it is our conventional choice to make use of this circumstance and set $\epsilon = 1/2$.

The conventionalist thesis is thus dependent on the notion that only certain local states of affairs, by virtue of being directly observable, are factual and objective. For the logical empiricists, the philosophical school of thought to which Reichenbach belonged, this was a very natural position to take. Broadly speaking, the logical empiricists defended the viewpoint that in the analysis of what scientific theories tell us a distinction must be made between *observation terms* and *theoretical terms.* The former define the objective empirical content of a theory, which can be formulated by referring to observable things, the latter serve first of all as mental tools that enable us to make predictions (again about observable things) – they are intermediaries that help us make connections between statements describing observable initial conditions and statements describing future states of affairs (predictions). These theoretical terms are introduced by us via definitions that link them to observation terms. Local facts, like the positions of the hands of a clock, are paradigmatic for what can be described by observation terms. In contrast, *simultaneity* is a theoretical notion that can be introduced in a variety of ways without doing harm to the empirical content of the theory. In fact, the logical empiricists were inspired by Einstein's presentation of special relativity and his introduction of simultaneity *by definition* and thought to capture the spirit of Einstein's example in their general analysis of scientific theories.

However, commenting on this with hindsight and from a modern point of view, we can safely say that it is far-fetched to suppose that Einstein was implicitly proposing a general scheme for an empiricist philosophy of science in his 1905 paper – we have already

commented on this before. Einstein's aim was obviously to convince his readers that the temporal structure assumed by classical mechanics is not sacrosanct – this is the key to special relativity. In order to make his point it was essential for Einstein to discuss how simultaneity can actually be established and to show that the most direct facts about which everyone agrees – the readings of the individual clocks – do not suffice to fix which events are simultaneous. This is something quite different from arguing that there is no objective simultaneity relation at all. Einstein's target was not the uniqueness of the synchronization rule in relativity theory (there is no mention of alternative synchronization procedures in Einstein's 1905 paper, nor in his later work), but rather the in his days universally assumed uniqueness of the *classical* conception of time. His goals in the paper were physical, not philosophical. In fact, as we have already pointed out, in later philosophical statements Einstein explicitly distanced himself from logical empiricist and operationalist claims.

More important than the question of what exactly were Einstein's own intentions and views is that the restriction of objectivity to directly observable states of affairs has long disappeared from the philosophy of science and the philosophy of physics. That something cannot be observed by humans in an immediate way would today no longer be accepted as conclusive evidence that it does not constitute an objective feature of reality, not even by empiricist philosophers. Present-day physics is, of course, full of not-directly observable entities, and these *theoretical entities* are at least *candidates* for being really existing things out there in the physical world. It would be beyond the scope of the present chapter to go into the debates surrounding scientific realism and empiricism in the philosophy of science, but it can safely be said that it is now generally accepted that there may be good reasons for accepting descriptions as candidates for being (approximately) true of reality, even if these descriptions partly refer to unobservable features.

In general, unobservability will diminish the epistemological warrant for accepting existence claims; for example, since we cannot directly observe subatomic particles, we are less certain about their existence than about the existence of tables and chairs. Similarly, the fact that we cannot observe simultaneity directly should give us some pause in making firm statements about its existence and character. In the case of simultaneity the lack of direct observability is not due to smallness of dimensions, but purely to the nonlocal character of the simultaneity relation. However, this nonlocality implies only a relatively mild lack of observability; after all, it *is* possible to go from one place to another and to list and compare all the directly observable local results we find on our way. It is, therefore, possible to say something, on the basis of direct local observations, about the relations between distant events. It is true that the observability of simultaneity itself may remain moot, but the question of whether all values of ϵ between 0 and 1 in Reichenbach's formula correspond to equally viable descriptions of nature can, nevertheless, be subjected to empirical investigation. Since the value of ϵ has implications for the symmetry properties of the description, considerations about the symmetries in the pattern of local observations appear to be especially relevant.

The original formulation of Einstein's synchronization procedure already shows that it makes use of a symmetry property of spacetime. It is not something certain *a priori* that the one-way velocity of light *can* be taken to be the same in all directions (in a given inertial frame of reference); but in fact this *is* possible without contradiction in Minkowski spacetime, in any inertial frame. The consistency of Einstein simultaneity thus expresses a physical fact about Minkowski spacetime: Minkowski spacetime is homogeneous and isotropic. Actually, in his 1905 paper Einstein uses this homogeneity and isotropy explicitly, as a premise in his derivation of the Lorentz transformations. Also Minkowski, in his seminal 1908 paper, pays explicit attention to the role of global symmetries in the formulation of relativity theory. He introduces standard coordinates x, y, z, and t in terms of which the mechanical and electromagnetic equations assume a preferred form (namely the standard one, which among other things makes the speed of light isotropic). He assigns these preferred coordinates a status like that of the privileged Cartesian coordinates in Euclidean geometry, which also latch onto a symmetry of the geometry. In his description of the procedure Minkowski explains that physical equations must be distilled from regularities in observed phenomena – although it is true that the building blocks of our knowledge come from local observations, the regularities only become visible if we compare, relate, and order these local data. Global aspects of the situation are, therefore, automatically relevant, and in particular the isotropy and homogeneity of spacetime stand out through the possibility of giving the physical laws completely symmetrical forms. Even Reichenbach, as we have already noted, considers this isotropy and homogeneity as objective and factual since it is an empirical result that a unique consistent description *is possible* in which the laws display identi-

cal properties at all points in space and time, and in all directions (namely, the description with $\epsilon = 1/2$). We may add that this existence claim is actually false in most general relativistic spacetimes, so that its truth in special relativity tells us something physically distinctive about Minkowski spacetime.

This emphasis on the relation between simultaneity and spacetime symmetry is in agreement with *Malament's* well-known result [6.31] that Einstein simultaneity is the only nontrivial equivalence relation that can be defined (in the mathematical sense) from the relation of causal connectibility. Malament demonstrates two things: first, given an inertial worldline and a spacetime point on it, the hyperplane that is Minkowski-orthogonal to the worldline through the given point on it can be constructed using null lines (light signals) – this construction is similar to the construction of the normal to a line in one of its points in Euclidean geometry. Second, he proves a uniqueness result: this orthogonal hyperplane is the only locus of points that does not coincide with the whole of Minkowski spacetime, implements an equivalence relation, and is invariant under all causal automorphisms (mappings of Minkowski spacetime in itself that preserve the causal structure) that leave the given worldline invariant. This latter result can be made plausible by visualization: think of four-dimensional rotations around the given worldline as an axis – only hyperplanes orthogonal to the worldline will be transformed into themselves by these transformations.

Now, if the causal theory of time is read as saying that a temporal relation is factual if and only if it can be defined in terms of the causal structure of Minkowski spacetime, Malament's theorem shows that Minkowski-orthogonality is the only nontrivial factual equivalence relation with respect to an inertial frame (represented by an inertial worldline, the *time axis* of the inertial system). If we accept that simultaneity should be an equivalence relation (i. e., symmetrical, reflexive, and transitive), it follows that $\epsilon = 1/2$ simultaneity (in Reichenbach's terminology) is factual and, moreover, unique (and, therefore, not conventional).

Reichenbach would not have accepted the conclusion that this result settles the conventionality debate, for he formulated his causal theory of time in terms of a constraint (*along a causal chain events cannot be simultaneous but must stand in the earlier–later relation – beyond this anything goes*) rather than committing himself to the position that simultaneity would be factual if it were a causally defined equivalence relation. He would, of course, admit that Minkowski *orthogonality* can be defined in the way we have sketched, but he would not automatically be forced to accept this Minkowski orthogonality as implementing simultaneity. In fact, *Reichenbach* explicitly denied one of the premises of Malament's proof, namely that simultaneity has to be an equivalence relation *with the same definition* in all directions [6.10]. This needs some explanation; of course, if events E_1 and E_2 are simultaneous, in this order, they also possess equal time coordinates in their reverse order, so that the relation is automatically symmetrical. However, Reichenbach did not require or accept that the *mathematical form* of the relation has to be the same in both directions. Indeed, in Reichenbach's formula $\epsilon \neq 1/2$ in one direction, from E_1 to E_2, implies $1 - \epsilon$, with $\epsilon \neq (1 - \epsilon)$, in the other direction so that the procedure of synchronizing must be different in the two directions. In fact, if Reichenbach had demanded that simultaneity is an equivalence relation *according to the same rule* in all directions, he would have immediately found $\epsilon = 1/2$.

Modern conventionalists respond to Malament's result in a similar way, by denying that simultaneity must be an equivalence relation; or they relax other conditions of Malament's proof in order to create room for alternative simultaneity relations [6.29, 32]. For example, if invariance under *all* causal automorphisms is no longer required, Malament's proof need not go through.

However, Malament's result does show uncontroversially that in Minkowski spacetime there *exists* a *candidate simultaneity* relation that is unique in being maximally adapted to the symmetry of the causal structure, namely the orthogonality relation with respect to the parallel inertial worldlines. This orthogonality relation is invariant under those symmetry transformations of Minkowski spacetime (elements of the Poincaré group) that leave invariant a given congruence of inertial worldlines. It is adapted to Minkowski geometry in a way that is very similar to the way Cartesian coordinate axes are adapted to the spatial Euclidean geometry. Its global existence is characteristic of the (flat) spacetime geometry of Minkowski spacetime, just as the global existence of Cartesian coordinates is characteristic of Euclidean geometry.

The debate about conventionality thus boils down to the question of whether or not using these privileged coordinates is conventional; their privileged status itself, of being adapted to the geometry and bringing out a symmetry in spacetime, is uncontested. Now it is certainly true that there is no compulsion for anyone to use Einstein simultaneity for the construction of time coordinates and that one may freely decide

to employ a conventionally chosen other time coordinate – the choice of coordinates is free, in relativity theory as well as in any other physical theory. There may even be particular contexts in which the use of nonstandard time coordinates recommends itself, for instance for the purposes of calculations. However, this seems a pragmatic point, of secondary importance in the discussion about the status of Einstein simultaneity as representing an objective physical feature of spacetime.

6.3.3 Simultaneity, Slow Clocks, and Conventionality in Noninertial Systems

In Sect. 6.2.3 we considered rotating frames of reference and encountered the Sagnac effect: there is a gap between the round trip arrival times of signals propagating along with the rotation and signals going in the opposite direction. As we have seen, this time gap is independent of the nature of the signals. So if we transport two clocks along a circle with radius r around the center of a rotating disk, one clockwise and one counterclockwise, while keeping their velocities constant with respect to their local comoving inertial frames, there will be a difference $\Delta t = 4\pi\omega r^2/(c^2 - \omega^2 r^2)$ between their return times (measured in the laboratory time t). Now, it is well known that the indications of traveling clocks conform to standard simultaneity in an inertial frame in the limiting case in which the clocks move very slowly (with respect to the given inertial frame). This, in fact, is one way in which the physical significance of Einstein simultaneity shows itself [6.33]: Einstein simultaneity approximates classical simultaneity in the classical limit. That is, in the classical limit of processes in which only very low velocities occur (relative to the velocity of light) within an inertial system, classical mechanics will become applicable in very good approximation, and in the equations Einstein simultaneity will coalesce with the classical simultaneity relation. It has to be noted that this will be so in *every* inertial system, each one with its own relativistic simultaneity, so that here there is no question of going back to the full classical time structure in a classical limit. Still, *per* inertial system, Einstein simultaneity is the natural generalization of classical simultaneity and in the limit of very low velocities it relates to physical processes formally in the same way Newtonian simultaneity relates to them in classical mechanics – in particular, very slowly moving clocks remain synchronized with resting laboratory clocks.

Taking our minds back to the rotating disc, we see that if the two clocks are transported very slowly with respect to the disc, they will remain synchronized according to standard simultaneity in their comoving local inertial frames. The Sagnac effect now tells us that this slow clock transport cannot be used to define an unambiguous global time coordinate on the rotating disc; the result will depend on whether a clockwise or counterclockwise path is chosen. In general, the result of synchronization by slow clock transport will be path dependent. Neither the Einstein light signal procedure, nor the slow transport of clocks can, therefore, be used to establish a global notion of simultaneity on the rotating disc. This is a result that can be generalized and is typical of accelerated frames of reference.

6.3.4 Simultaneity, Symmetry, and Time Flow

So again, we have found that global Einstein simultaneity reflects the symmetry embodied in inertial systems; in inertial systems this $\epsilon = 1/2$ simultaneity allows a simple, adapted, formulation of the laws, conforms to slow clock transport and many other almost-classical processes, and agrees with global Minkowski-orthogonality with respect to the worldlines representing the state of rest. In noninertial frames all such arguments apply only locally. The rotating system illustrates the situation very well: in each point on the disc standard simultaneity can be established just as in an inertial system, but these local simultaneities do not combine into a physically meaningful global time coordinate.

Summing up, Einstein simultaneity has a special status because it leads to global foliations of Minkowski spacetime that respect the physical symmetries. Inertial reference frames, characterized by congruences of parallel straight worldlines, bring out the homogeneity and isotropy that is present in Minkowski spacetime if they are equipped with Einstein simultaneity. In other words, in inertial systems Einstein simultaneity is clearly a physically significant global relation, representing a physical fact.

However, this physical significance relates to symmetry properties of spacetime rather than to a *causal* interconnection; as we have noted before, the relativistic physical evolution equations do not need simultaneity for their formulation, because all interactions are local. Slicing up spacetime according to the Einstein simultaneity associated with an inertial system does not correspond to a grouping together of events that belong together in a way that is causally or dynami-

cally significant. Moreover, the relativity and, therefore, nonuniqueness of simultaneity precludes an interpretation of Einstein simultaneity hyperplanes in terms of a progressing universal now. This break between intuitive features of simultaneity (according to everyday intuition we are in direct causal contact with simultaneous events at a distance and time flows because the universal now shifts to the future) and the significance of relativistic simultaneity probably constitute a major part of the motivation behind the doctrine that simultaneity in relativity is conventional. Indeed, an important connection between simultaneity and intuitions about time is severed once Einstein simultaneity is accepted as the simultaneity relation in Minkowski spacetime, and this may seem to make the notion of relativistic simultaneity insignificant and conventional. However, as we have seen, there are nevertheless good reasons to assign Einstein simultaneity a factual status in relativity theory.

In noninertial, accelerating frames of reference the spacetime symmetry of Minkowski spacetime will, in general, be broken or rather masked, because the defining congruences of worldlines will not be parallel. In such cases it cannot be expected that a physically significant global simultaneity relation, adapted to the frames in question, will exist. A global time *coordinate* can in these circumstances still be introduced, and simultaneity can be defined via sameness of value of this time *coordinate*. However, now this becomes a matter of pragmatic choice and, therefore, of convention. This is true in noninertial frames of reference, like the rotating disc, and also in generally relativistic spacetimes in which there are no global symmetries. These noninertial frames of reference and general relativistic spacetimes, rather than the inertial frames of special relativity, constitute the arena in which the thesis that distant relativistic simultaneity is conventional finds its natural habitat and justification.

References

6.1 I. Newton: Sir Isaac Newton's Mathematical Principles of Natural Philosophy and his System of the World, 2 Vols., ed. by F. Cajori (Univ. California Press, Berkeley 1962) transl. by A. Motte

6.2 L. Sklar: *Space, Time and Spacetime* (Univ. of California Press, Berkeley 1974)

6.3 A. Einstein: Zur Elektrodynamik bewegter Körper, Ann. Phys. **17**, 891–921 (1905)

6.4 J. Stachel, D.C. Cassidy, J. Renn, R. Schulmann (Eds.): *The Collected Papers of Albert Einstein*, Vol. 2 (Princeton Univ. Press, Princeton 1989)

6.5 H.A. Lorentz, A. Einstein, H. Minkowski, H. Weyl: *The Principle of Relativity* (Methuen, London 1923), reprinted 1952 (Dover, New York)

6.6 H. Minkowski: Raum und Zeit, Phys. Z. **10**, 104–111 (1909)

6.7 D. Howard: Einstein and the development of twentieth-century philosophy of science. In: *The Cambridge Companion to Einstein*, ed. by M. Janssen, C. Lehner (Cambridge Univ. Press, Cambridge 2014)

6.8 D. Dieks: The adolescence of relativity: Einstein, Minkowski, and the philosophy of space and time. In: *Minkowski Spacetime: A Hundred Years Later*, ed. by V. Petkov (Springer, Dordrecht 2010) pp. 225–245

6.9 D. Dieks: Reichenbach and the conventionality of distant simultaneity in perspective. In: *The Present Situation in the Philosophy of Science*, ed. by F. Stadler (Springer, Dordrecht 2010) pp. 315–333

6.10 H. Reichenbach: *The Philosophy of Space and Time* (Dover, New York 1957)

6.11 P.W. Bridgman: Einstein's theories and the operational point of view. In: *Albert Einstein: Philosopher-Scientist*, ed. by P.A. Schilpp (Open Court, La Salle 1949) pp. 333–355

6.12 A. Einstein: Remarks to the essays appearing in this collective volume. In: *Albert Einstein: Philosopher-Scientist*, ed. by P.A. Schilpp (Open Court, La Salle 1949) pp. 665–688

6.13 D. Dieks: Space, time and coordinates in a rotating world. In: *Relativity in Rotating Frames*, ed. by G. Rizzi, M.L. Ruggiero (Kluwer, Dordrecht 2004) pp. 29–42

6.14 B. Dainton: *Time and Space*, 2nd edn. (Acumen, Chesham 2010)

6.15 J. Faye: *The Reality of the Future* (Odense Univ. Press, Odense 1989)

6.16 S. Savitt: Being and becoming in modern physics. In: The Stanford Encyclopedia of Philosophy, ed. by E.N. Zalta (Fall 2013 Edition) available on line at http://plato.stanford.edu/archives/fall2013/entries/spacetime-bebecome/

6.17 T. Sider: *Four-Dimensionalism: An Ontology of Persistence and Time* (Oxford Univ. Press, Oxford 2001)

6.18 N. Markosian: *Studies in Metaphysics Ser*, A Defense of Presentism, Vol. 1 (Oxford Univ. Press, Oxford 2003)

6.19 C.W. Rietdijk: A rigorous proof of determinism derived from the special theory of relativity, Philos. Sci. **33**, 341–344 (1966)

6.20 C.D. Broad: *Scientific Thought* (Kegan Paul, London 1923)

6.21 J.M.E. McTaggart: The unreality of time, Mind **68**, 457–484 (1908)
6.22 H. Putnam: Time and physical geometry, J. Philos. **64**, 240–247 (1967), reprinted in *Putnam's Collected Papers*, Vol. 1 (Cambridge Univ. Press, Cambridge 1975)
6.23 D. Dieks: The physics and metaphysics of time, Eur. J. Anal. Philos. **8**, 103–120 (2012)
6.24 W.A. Craig: *Time and the Metaphysics of Relativity* (Kluwer Academic, Dordrecht 2001)
6.25 C. Bourne: *A Future for Presentism* (Oxford Univ. Press, Oxford 2006)
6.26 Y. Balashov, M. Janssen: Presentism and relativity, Br. J. Philos. Sci. **54**, 327–346 (2003)
6.27 D. Dieks: Becoming, relativity and locality. In: *The Ontology of Spacetime*, ed. by D. Dieks (Elsevier, Amsterdam 2006) pp. 157–176
6.28 D. Dieks: Special relativity and the flow of time, Philos. Sci. **55**, 456–460 (1988)
6.29 A. Janis: Conventionality of simultaneity. In: Stanford Encyclopedia of Philosophy, ed. by E.N. Zalta (Fall 2010 Edition) available online at http://plato.stanford.edu/archives/fall2010/entries/spacetime-convensimul/
6.30 M. Jammer: *Concepts of Simultaneity* (Johns Hopkins Univ. Press, Baltimore 2006)
6.31 D. Malament: Causal theories of time and the conventionality of simultaneity, Noûs **11**, 293–300 (1997)
6.32 S. Sarkar, J. Stachel: Did Malament prove the non-conventionality of simultaneity in the special theory of relativity?, Philos. Sci. **66**, 208–220 (1999)
6.33 B. Ellis, P. Bowman: Conventionality in distant simultaneity, Philos. Sci. **34**, 116–136 (1967)

Part B

Foundati

Part B Foundational Issues

7. Rigid Motion and Adapted Frames

Stephen N. Lyle

The aim here is to describe the rigid motion of a continuous medium in special and general relativity. Section 7.1 defines a rigid rod in special relativity, and Sect. 7.2 shows the link with the space coordinates of a certain kind of accelerating frame in flat spacetimes. Section 7.3 then sets up a notation for describing the arbitrary smooth motion of a continuous medium in general curved spacetimes, defining the proper metric of such a medium. Section 7.4 singles out rigid motions and shows that the rod in Sect. 7.1 undergoes rigid motion in the more generally defined sense. Section 7.5 defines a rate of strain tensor for a continuous medium in general relativity and reformulates the rigidity criterion. Section 7.6 aims to classify all possible rigid motions in special relativity, reemphasizing the link with semi-Euclidean frames adapted to accelerating observers in special relativity. Then, Sects. 7.7 and 7.8 describe rigid motion without rotation and rigid rotation, respectively. Along the way we introduce the notion of Fermi–Walker transport and discuss its relevance for rigid motions. Section 7.9 brings together all the above themes in an account of a recent generalization of the notion of uniform acceleration, thereby characterizing a wide class of rigid motions.

7.1 Rigid Rod in Special Relativity

Of course, we know what happens to a rigid rod when it has uniform motion relative to an inertial frame $\mathcal{I}$. In other words, we know what we want rigidity to mean in that context. But can we say how a rigid rod should behave when it accelerates? Can we still have some kind of rigidity?

Let A and B be the left- and right-hand ends of the rod, respectively, and consider motion $x_A(t)$ and $x_B(t)$ along the axis from A to B (Fig. 7.1). Let us first label the particles in the rod by their distance s to the right of A when the system is stationary in some inertial frame (Fig. 7.2). This idea of labeling particles will prove extremely useful when considering continuous media later on. In this case, we imagine the rod as a strictly one-dimensional, continuous row of particles.

Now let A have motion $x_A(t)$ relative to an inertial frame $\mathcal{I}$ (Fig. 7.3) and let $X(s,t)$ be a function giving the position of particle s at time t as

$$x_s(t) = x_A(t) + X(s,t) ,$$

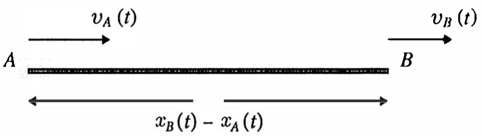


Fig. 7.1 Material rod in motion along its axis in an inertial frame $\mathcal{I}$. The position of the left-hand end A is given by $x_A(t)$ at time t, and the position of the right-hand end B is given by $x_B(t)$

where in fact we require

$$X(0,t)=0\,,\quad X(D,t)=x_B(t)-x_A(t)\,.$$

Let us require the element between s and $s+\delta s$ to have coordinate length

$$\left[1-\frac{v(s,t)^2}{c^2}\right]^{1/2}\delta s\,, \tag{7.1}$$

where $v(s,t)$ is its instantaneous coordinate velocity, with $v(0,t)=v_A(t)$. This is precisely the criterion suggested by *Rindler* [7.1, pp. 39–40]. We can integrate to find

$$X(s,t)=\int_0^s\left[1-\frac{v(s',t)^2}{c^2}\right]^{1/2}\mathrm{d}s'\,. \tag{7.2}$$

This implies that

$$X_B=\int_0^D\left[1-\frac{v(s',t)^2}{c^2}\right]^{1/2}\mathrm{d}s'\,. \tag{7.3}$$

Note the highly complex equation this gives for the speed function $v(s,t)$, namely,

$$v(s,t)=v_A(t)+\frac{\partial X(s,t)}{\partial t}\,. \tag{7.4}$$

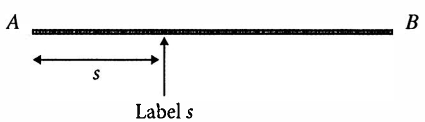


Fig. 7.2 Stationary material rod in an inertial frame $\mathcal{I}$. Labeling the particles in the rod by their distance s from A, so that $s\in[0,D]$

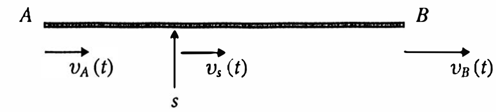


Fig. 7.3 Material rod with arbitrary motion in an inertial frame $\mathcal{I}$

Let us observe carefully that we are not assuming any simple Galilean addition law for velocities here. This is a straightforward differentiation with respect to t of the formula for the coordinate position of atom s at time t, namely, $x_A(t)+X(s,t)$. The partial time derivative of X is not the velocity of s relative to A, that is, it is not the velocity of s measured in a frame moving with A.

Now (7.2) seems to embody the idea of the rod being rigid. In fact this rod could no longer be elastic, in the sense that (7.1) only allows the element δs to relativistically contract for the value of its instantaneous speed, forbidding any other contortions. One could well imagine the rod undergoing a very complex deformation along its length, in which relativistic contraction effects were quite negligible compared with a certain looseness in the molecular bonding, but we are not talking about this. In fact we are seeking a definition of rigidity that does not refer to the microscopic structure.

7.1.1 Equation of Motion for Points on the Rod

So far the main equations for the atom labeled s on the rod are (7.2) and (7.4), namely,

$$X(s,t)=\int_0^s\left[1-\frac{v(s',t)^2}{c^2}\right]^{1/2}\mathrm{d}s' \tag{7.5}$$

and

$$v(s,t)=v_A(t)+\frac{\partial X(s,t)}{\partial t}\,. \tag{7.6}$$

The first implies that

$$\frac{\partial X(s,t)}{\partial s}=\left[1-\frac{v(s,t)^2}{c^2}\right]^{1/2}\,. \tag{7.7}$$

We can write one nonlinear partial differential equation for $X(s,t)$ by eliminating $v(s,t)$ to give

$$c^2\left(\frac{\partial X}{\partial s}\right)^2+\left[\frac{\partial X}{\partial t}+v_A(t)\right]^2=c^2\,. \tag{7.8}$$

This is effectively the equation that we have to solve to find the length of our rod. It is important to see that there is a boundary condition too, namely,

$$0 = \left.\frac{\partial X(s,t)}{\partial t}\right|_{s=0} , \tag{7.9}$$

because we do require $v(0,t) = v_A(t)$ in conjunction with (7.6).

We shall find a solution to this problem, although not by solving (7.8) directly. Instead we shall follow a circuitous but instructive route and end up guessing the relevant solution.

7.2 Frame for an Accelerating Observer

We remain for the moment in the context of special relativity. Let AO be the name for an observer moving with the left-hand end A of the proposed rod. AO is an accelerating observer and it is well known [7.2] that such a person can find well-adapted coordinates y^μ with the following properties (where the Latin index runs over $\{1,2,3\}$):

- First of all, any curve with all three y^i constant is timelike and any curve with y^0 constant is spacelike.
- At any point along the worldline of AO, the zero coordinate y^0 equals the proper time along that worldline.
- At each point of the worldline of AO, curves with constant y^0 which intersect it are orthogonal to it where they intersect it.
- The metric has the Minkowski form along the worldline of AO.
- The coordinates y^i are Cartesian on every hypersurface of constant y^0.
- The equation for the worldline of AO has the form $y^i = 0$ for $i = 1, 2, 3$.

Such coordinates could be called semi-Euclidean (SE). They are often called Rindler coordinates.

Let us consider a one-dimensional (1-D) acceleration and temporarily drop the subscript A on the functions $x_A(t)$ and $v_A(t)$ describing the motion of AO in the inertial frame $\mathcal{I}$. The worldline of the accelerating observer is given in inertial coordinates by

$$t = \sigma , \quad x = x(\sigma) , \quad \frac{\mathrm{d}x}{\mathrm{d}\sigma} = v(\sigma) , \tag{7.10}$$

$$\frac{\mathrm{d}^2x}{\mathrm{d}\sigma^2} = a(\sigma) , \quad y(\sigma) = 0 = z(\sigma) , \tag{7.11}$$

using the time t in $\mathcal{I}$ to parameterize. The proper time $\tau(\sigma)$ of AO is given by

$$\frac{\mathrm{d}\tau}{\mathrm{d}\sigma} = \left(1 - \frac{v^2}{c^2}\right)^{1/2} . \tag{7.12}$$

The coordinates y^μ are constructed on an open neighborhood of the AO worldline as follows (Fig. 7.4). For an event (t, x, y, z) not too far from the worldline, there is a unique value of τ and hence also the parameter σ such that the point lies in the hyperplane of simultaneity (HOS) of AO when its proper time is τ. This HOS is given by

$$t - \sigma(\tau) = \frac{v(\sigma(\tau))}{c^2} \left[x - x(\sigma(\tau))\right] , \tag{7.13}$$

which solves, for any x and t, to give $\sigma(\tau)(x,t)$.

The semi-Euclidean coordinates attributed to the event (t, x, y, z) are, for the time coordinate y^0, (c times) the proper time τ found from (7.13) and, for the spatial coordinates, the spatial coordinates of this event in an instantaneously comoving inertial frame (ICIF) at proper time τ of AO. In fact, every other event in this instantaneously comoving inertial frame is attributed to the same time coordinate $y^0 = c\tau$ and the appropriate spatial coordinates borrowed from this frame. Of course, the HOS of AO at time τ is also the one borrowed from the ICIF.

There is just one detail to get out of the way: there are many different ICIFs for a given τ, and there are even many different ways to choose these frames as a smooth function of τ as one moves along the AO worldline, rotating back and forth around various axes in the original inertial frame $\mathcal{I}$ as τ progresses. For the present purposes, we choose a sequence with no rotation about any space axis in the instantaneous local rest frame. It can always be done by solving the Fermi–Walker (FW) transport equations (Sect. 7.7). The semi-Euclidean coordinates are then given by

$$\begin{cases} y^0 = c\tau , \\ y^1 = \dfrac{[x - x(\sigma)] - v(\sigma)(t-\sigma)}{\sqrt{1 - \frac{v^2}{c^2}}} , \\ y^2 = y , \\ y^3 = z , \end{cases} \tag{7.14}$$

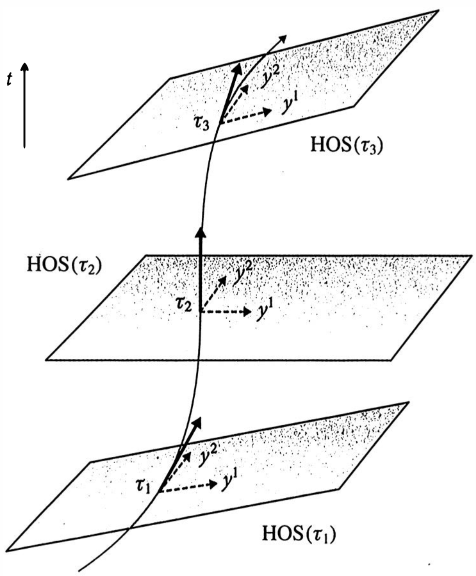


Fig. 7.4 Constructing an SE frame for an accelerating observer. View from an inertial frame with time coordinate t. The curve is the observer worldline given by (7.10). Three HOS are shown at three successive proper times τ_1, τ_2, and τ_3 of the observer. These hyperplanes of simultaneity are borrowed from the instantaneously comoving inertial observer, as are the coordinates y^1, y^2, and y^3 used to coordinatize them. Only two of the latter coordinates can be shown in the spacetime diagram

where $\sigma = \sigma(t, x)$ as found from (7.13). The inverse transformation, from semi-Euclidean coordinates to inertial coordinates, is given by

$$\begin{cases} t = \sigma(y^0) + \dfrac{v(y^0)}{c^2} y^1 \left[1 - \dfrac{v(y^0)^2}{c^2}\right]^{-1/2} , \\ x = x(y^0) + y^1 \left[1 - \dfrac{v(y^0)^2}{c^2}\right]^{-1/2} , \\ y = y^2 , \\ z = y^3 , \end{cases} \tag{7.15}$$

where the function $\sigma(y^0)$ is just the expression relating inertial time to proper time for the accelerating observer, and the functions $x(y^0)$ and $v(y^0)$ should really be written $x(\sigma(y^0))$ and $v(\sigma(y^0))$, respectively.

The above relations are not very enlightening. They are only displayed to show that the idea of such coordinates can be made perfectly concrete. One calculates the metric components in this frame, namely,

$$g_{00} = \frac{1}{g^{00}} = \left[1 + \frac{\frac{a(\sigma)[x - x(\sigma)]}{c^2}}{1 - \frac{v(\sigma)^2}{c^2}}\right]^2 , \tag{7.16}$$

where $\sigma = \sigma(t, x)$ as found from (7.13), and

$$g_{i0} = 0 = g_{0i} , \quad g_{ij} = -\delta_{ij} , \quad i, j \in \{1, 2, 3\} , \tag{7.17}$$

and checks the list of requirements for the coordinates to be suitably adapted to the accelerating observer.

Although perfectly concrete, the coordinates are not perfectly explicit: the component g_{00} of the semi-Euclidean metric has been expressed in terms of the original inertial coordinates! This can be remedied as follows. One observes that, with the help of (7.13),

$$y^1 = [x - x(\sigma)]\sqrt{1 - \frac{v^2}{c^2}} . \tag{7.18}$$

One calculates the 4-acceleration in the inertial frame to be

$$a^\mu := \frac{d^2 x^\mu}{d\tau^2} = a\left(1 - \frac{v^2}{c^2}\right)^{-2} \left(\frac{v}{c}, 1, 0, 0\right) , \tag{7.19}$$

and transforms this by Lorentz transformation to the inertial frame instantaneously comoving with the observer to find only one nonzero 4-acceleration component in that frame, which is called the absolute acceleration of the observer

$$a_{01} := \text{absolute acceleration} = a\left(1 - \frac{v^2}{c^2}\right)^{-3/2} . \tag{7.20}$$

The notation a_{01} for the 1-component of the absolute acceleration will appear again in (7.52). One now has the more comforting formula

$$g_{00} = \frac{1}{g^{00}} = \left[1 + \frac{a_{01}(\sigma) y^1}{c^2}\right]^2 . \tag{7.21}$$

To obtain still more explicit formulas, one needs to consider a specific motion $x(\sigma)$ of AO, the classic example being uniform acceleration

$$x(\sigma) = \frac{c^2}{g}\left[\left(1 + \frac{g^2\sigma^2}{c^2}\right)^{1/2} - 1\right] , \quad t = \sigma , \tag{7.22}$$

where g is some constant with units of acceleration. This does not look like a constant acceleration in the inertial frame

$$\frac{\mathrm{d}x}{\mathrm{d}\sigma} = \frac{g\sigma}{\left(1+\frac{g^2\sigma^2}{c^2}\right)^{1/2}} , \qquad \frac{\mathrm{d}^2x}{\mathrm{d}\sigma^2} = \frac{g}{\left(1+\frac{g^2\sigma^2}{c^2}\right)^{3/2}} . \tag{7.23}$$

However, the 4-acceleration defined in the inertial frame $\mathcal{I}$ by

$$a^\mu = \frac{\mathrm{d}^2x^\mu}{\mathrm{d}\tau^2} , \tag{7.24}$$

where τ is the proper time, has constant magnitude. It turns out that

$$a^2 := a_\mu a^\mu = -g^2 ,$$

with a suitable convention for the signature of the metric.

In this case, the transformation from inertial to semi-Euclidean coordinates is

$$y^0 = \frac{c^2}{g}\tanh^{-1}\frac{ct}{x+\frac{c^2}{g}} , \tag{7.25}$$

$$y^1 = \left[\left(x+\frac{c^2}{g}\right)^2 - c^2t^2\right]^{1/2} - \frac{c^2}{g} , \tag{7.26}$$

$$y^2 = y , \quad y^3 = z ,$$

and the inverse transformation is

$$t = \frac{c}{g}\sinh\frac{gy^0}{c^2} + \frac{y^1}{c}\sinh\frac{gy^0}{c^2} , \tag{7.27}$$

$$x = \frac{c^2}{g}\left(\cosh\frac{gy^0}{c^2} - 1\right) + y^1\cosh\frac{gy^0}{c^2} , \tag{7.28}$$

$$y = y^2 , \quad z = y^3 .$$

One finds the metric components to be

$$g_{00} = \left(1+\frac{gy^1}{c^2}\right)^2 , \quad g_{0i} = 0 = g_{i0} , \qquad g_{ij} = -\delta_{ij} , \tag{7.29}$$

for $i,j \in \{1,2,3\}$, in the semi-Euclidean frame. Interestingly, this metric is static, i. e., g_{00} is independent of y^0. It is the only semi-Euclidean metric that is [7.3].

It is worth pausing to wonder why AO should adopt such coordinates. It must be comforting to attribute one's own proper time to events that appear simultaneous. But what events are simultaneous with AO? In the above construction, AO borrows the hyperplane of simultaneity of an inertially moving observer, who does not have the same motion at all. AO also borrows the lengths of this inertially moving observer.

But if AO were carrying a rigid measuring rod, what lengths would be measured with it? In fact the rigid rod of Sect. 7.1 measures the spatial coordinates of AO when this observer uses semi-Euclidean coordinates. Let us prove this for the case of a uniform acceleration g, where formulas are explicit.

We write down the path of a point with some fixed spatial coordinate s along the axis of acceleration (putting the other spatial coordinates equal to zero). The formula we have for the path of the origin of the SE frame as expressed in Minkowski coordinates is

$$x_A(t) = \frac{c^2}{g}\left(\sqrt{1+\frac{g^2t^2}{c^2}} - 1\right) , \tag{7.30}$$

giving a coordinate velocity

$$v_A(t) = \frac{gt}{\sqrt{1+\frac{g^2t^2}{c^2}}} . \tag{7.31}$$

The formula for the path of the point at fixed SE spatial coordinate s from the origin as expressed in Minkowski coordinates is

$$x_s(t) = X(s,t) + x_A(t) = \frac{c^2}{g}\left[\sqrt{\left(1+\frac{gs}{c^2}\right)^2 + \frac{g^2t^2}{c^2}} - 1\right] . \tag{7.32}$$

We are going to show that the function $X(s,t)$ defined by the last relation actually satisfies our equation of motion (7.8) in the case where the function $x_A(t)$ gives the path of the left-hand end A of the rod, i. e., when the point A is uniformly accelerated by g.

Proof that (7.32) is a solution for (7.8): We begin with the partial derivatives

$$\frac{\partial X}{\partial t} = \frac{gt}{\sqrt{\left(1+\frac{gs}{c^2}\right)^2 + \frac{g^2t^2}{c^2}}} - v_A(t) , \tag{7.33}$$

$$\frac{\partial X}{\partial s} = \frac{1+gs/c^2}{\sqrt{\left(1+\frac{gs}{c^2}\right)^2 + \frac{g^2t^2}{c^2}}} . \tag{7.34}$$

Hence,

$$\left[\frac{\partial X}{\partial t}+v_A(t)\right]^2=\frac{g^2t^2}{\left(1+\frac{gs}{c^2}\right)^2+\frac{g^2t^2}{c^2}} \tag{7.35}$$

and

$$c^2\left(\frac{\partial X}{\partial s}\right)^2=\frac{c^2\left(1+\frac{gs}{c^2}\right)^2}{\left(1+\frac{gs}{c^2}\right)^2+\frac{g^2t^2}{c^2}}\,. \tag{7.36}$$

Adding the last two equations together, it is clear that we just get c^2, as required by (7.8). The boundary condition (7.9) is obviously satisfied too. ■

For a rod with arbitrary 1-D acceleration, the formulas are much more involved, due to the lack of explicitness, but the proof is nevertheless straightforward. So not only have we found the length of our rigid rod when it is accelerating along its own axis, but also we discover that any AO with 1-D motion could use it to measure semi-Euclidean coordinates along the direction of acceleration. This means that the rigid rod automatically satisfies what is sometimes called the ruler hypothesis, namely, it is at any instant of time ready to measure lengths in an instantaneously comoving inertial frame, since this is precisely the length system used by the semi-Euclidean coordinates.

The accelerating observer would not necessarily have to be holding one end of the rod. It could be lying with one end held fixed at some semi-Euclidean coordinate value $y^1=s_1$ and the other end would then remain at a constant coordinate value $y^1=s_2>s_1$. This is shown by exactly the same kind of analysis as above. In other words, if the rod always manages to occupy precisely this interval on the axis of the SE coordinate system, its length as viewed in the original inertial frame $\mathcal{I}$ will satisfy the rigidity equation (7.8). Hence, a rigid rod whose left-hand end is compelled to follow the worldline $y^1=s_1$ will always appear to have the same length s_2-s_1 to the SE observer.

7.3 General Motion of a Continuous Medium

The component particles of the medium are labeled by three parameters ξ^i, $i=1,2,3$, and the worldline of particle ξ is given by four functions $x^\mu(\xi,\tau)$, $\mu=0,1,2,3$, where τ is its proper time. In general relativity, the x^μ may be arbitrary coordinates in curved spacetime.

If $\xi^i+\delta\xi^i$ are the labels of a neighboring particle, its worldline is given by the functions

$$x^\mu(\xi+\delta\xi,\tau)=x^\mu(\xi,\tau)+x^\mu{}_{,i}(\xi,\tau)\delta\xi^i\,,$$

where the comma followed by a Latin index denotes partial differentiation with respect to the corresponding ξ. Note that the quantity $x^\mu{}_{,i}(\xi,\tau)\delta\xi^i$, representing the difference between the two sets of worldline functions, is formally a 4-vector, being basically an infinitesimal coordinate difference. However, it is not generally orthogonal to the worldline of ξ. In other words, it does not lie in the HOS of either particle.

To get such a vector we apply the projection tensor onto the instantaneous HOS. In inertial coordinates in a flat spacetime,

$$P^{\mu\nu}=\eta^{\mu\nu}-\dot{x}^\mu\dot{x}^\nu\,,$$

where $\eta^{\mu\nu}$ is the Minkowski metric and the dot denotes partial differentiation with respect to τ. In general relativity, the projection tensor takes the form

$$P^{\mu\nu}=g^{\mu\nu}-\dot{x}^\mu\dot{x}^\nu\,,$$

with $g^{\mu\nu}$ the metric tensor of the curved spacetime. The result is

$$\begin{aligned}\delta x^\mu &:= P^\mu{}_\nu x^\nu{}_{,i}(\xi,\tau)\delta\xi^i\\ &= x^\mu{}_{,i}\delta\xi^i-\dot{x}^\mu\dot{x}_\nu x^\nu{}_{,i}\delta\xi^i\,.\end{aligned} \tag{7.37}$$

One finds that application of the projection tensor corresponds to a simple proper-time shift of amount

$$\delta\tau=-g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu{}_{,i}\delta\xi^i\,,$$

so that

$$\delta x^\mu=x^\mu(\xi+\delta\xi,\tau+\delta\tau)-x^\mu(\xi,\tau)\,.$$

Indeed,

$$x^\mu(\xi+\delta\xi,\tau+\delta\tau)=x^\mu(\xi,\tau)+x^\mu{}_{,i}\delta\xi^i+\dot{x}^\mu\delta\tau\,,$$

and feeding in the proposed expression for $\delta\tau$, we do obtain precisely δx^μ as defined above, since

$$\begin{aligned}\delta x^\mu &= x^\mu(\xi,\tau)+x^\mu{}_{,i}\delta\xi^i+\dot{x}^\mu\delta\tau-x^\mu(\xi,\tau)\\ &= x^\mu{}_{,i}\delta\xi^i-g_{\nu\sigma}\dot{x}^\nu\dot{x}^\sigma{}_{,i}\delta\xi^i\dot{x}^\mu\,.\end{aligned}$$

What can we conclude from this analysis? The two particles ξ and $\xi + \delta\xi$ appear, in the instantaneous rest frame of either, to be separated by a distance δs given by

$$(\delta s)^2 = -\delta x \cdot \delta x = -\gamma_{ij}\delta\xi^i\delta\xi^j \,, \tag{7.38}$$

where

$$\gamma_{ij} = P_{\mu\nu}x^{\mu}{}_{,i}x^{\nu}{}_{,j} \,. \tag{7.39}$$

The quantity γ_{ij} is called the proper metric of the medium [7.4].

7.4 Rigid Motion of a Continuous Medium

At this point, one can introduce a notion of rigidity. One says that the medium undergoes rigid motion if and only if its proper metric is independent of τ. This is therefore expressed by

$$\dot{\gamma}_{ij} = 0 \,. \tag{7.40}$$

Under rigid motion the instantaneous separation distance between any pair of neighboring particles is constant in time, as they would see it. Note that this criterion is independent of the coordinates used because γ_{ij} is a scalar under coordinate transformation.

Let us see whether this coincides with the notion of rigidity discussed earlier, i. e., whether the rigid rod of Sect. 7.1 is rigid according to the new criterion, or put differently, whether the rod described in Sect. 7.1 is undergoing rigid motion according to the criterion (7.40). The labels ξ correspond to s in Sect. 7.1 (Fig. 7.2). In a given inertial frame, particle s has motion described by $X(s, t)$, where

$$\frac{\partial X}{\partial s} = \frac{1}{\gamma} \,, \qquad \gamma = \gamma\left(v(s,t)\right) \,,$$

and

$$v(s,t) = v_A(t) + \frac{\partial X}{\partial t} \,,$$

where $v_A(t)$ is the speed of the end of the rod. Suppose we now change to a frame moving instantaneously at speed $v(s, t)$ and measure the distance between particle s and particle $s + \delta s$ as viewed in this frame. Will it be constant in this model, as required for rigid motion? In the original frame where both particles are moving, we have separation

$$X(s+\delta s, t) - X(s,t) = \frac{\partial X}{\partial s}\delta s = \frac{\delta s}{\gamma} \,.$$

In the new frame moving at speed $v(s, t)$, this has length

$$\gamma\frac{\delta s}{\gamma} = \delta s = \text{constant} \,.$$

This is what rigid motion requires.

7.5 Rate of Strain Tensor

The aim here is to express the rigid motion condition $\dot{\gamma}_{ij} = 0$ in terms of derivatives with respect to the coordinates x^{μ} by introducing the relativistic analog of the rate of strain tensor in ordinary continuum mechanics.

The nonrelativistic strain tensor can be defined by

$$e_{ij} := \frac{1}{2}\left(\frac{\partial u_j}{\partial x_i} + \frac{\partial u_i}{\partial x_j}\right) \,,$$

where $u_i(x)$ are the components of the displacement vector of the medium, describing the motion of the point originally at x when the material is deformed. One also defines the antisymmetric tensor

$$\omega_{ij} := \frac{1}{2}\left(\frac{\partial u_j}{\partial x_i} - \frac{\partial u_i}{\partial x_j}\right) \,,$$

which describes the rotation occurring when the material is deformed. Clearly,

$$e_{ij} - \omega_{ij} = \frac{\partial u_i}{\partial x_j} \,,$$

and hence, if all distortions are small,

$$\Delta u_i = (e_{ij} - \omega_{ij})\Delta x_j \,.$$

We can consider that e_{ij} describes nonrotational distortions, i. e., stretching, compression, and shear.

In the present discussion, u_i is replaced by a velocity field v_i and we have a rate of strain tensor. The nonrelativistic rate of strain tensor is

$$r_{ij} = v_{i,j} + v_{j,i} \,, \tag{7.41}$$

where v_i is the 3-velocity field and the differentiation is with respect to ordinary Cartesian coordinates. Let us look for a moment at this tensor. The nonrelativistic condition for rigid motion is

$$r_{ij} = 0 \quad \text{everywhere} \,.$$

This equation implies

$$0 = r_{ij,k} = v_{i,jk} + v_{j,ik} \,, \tag{7.42}$$

$$0 = r_{jk,i} = v_{j,ki} + v_{k,ji} \,. \tag{7.43}$$

Subtracting (7.43) from (7.42) and commuting the partial derivatives, we find

$$v_{i,jk} - v_{k,ji} = 0 \,, \tag{7.44}$$

which, upon permutation of the indices j and k, yields also

$$v_{i,kj} - v_{j,ki} = 0 \,. \tag{7.45}$$

Adding (7.42) and (7.45), we obtain

$$v_{i,jk} = 0 \,,$$

which has the general solution

$$v_i = -\omega_{ij} x_j + \beta_i \,, \tag{7.46}$$

where ω_{ij} and β_i are functions of time only. The condition $r_{ij} = 0$ constrains ω_{ij} to be antisymmetric, i. e.,

$$\omega_{ij} = -\omega_{ji} \,,$$

and nonrelativistic rigid motion is seen to be, at each instant, a uniform rotation with angular velocity

$$\omega_i = \frac{1}{2} \varepsilon_{ijk} \omega_{jk}$$

about the coordinate origin, superimposed upon a uniform translation with velocity β_i. Because the coordinate origin may be located arbitrarily at each instant, rigid motion may alternatively be described as one in which an arbitrary particle in the medium moves in an arbitrary way while at the same time the medium as a whole rotates about this point in an arbitrary (but uniform) way. Such a motion has six degrees of freedom.

Note that when r_{ij} is zero, we can also deduce that $v_{i,i} = 0$, i. e., $\operatorname{div} v = 0$, which is the condition for an incompressible fluid. This is evidently a weaker condition than rigidity.

Let us see how this generalizes to special relativity. We return to the continuous medium in which particles are labeled by ξ^i, $i = 1, 2, 3$. Just as the coordinates x^μ are functions of the ξ^i and τ, so the ξ^i and τ can be regarded as functions of the x^μ, at least in the region of spacetime occupied by the medium. Following [7.4], we write

$$u^\mu := \dot{x}^\mu \,, \quad u^2 = 1 \,, \quad P_{\mu\nu} = \eta_{\mu\nu} - u_\mu u_\nu \,.$$

If f is an arbitrary function in the region occupied by the medium then

$$f_{,\mu} = f_{,i} \xi^i{}_{,\mu} + \dot{f} \tau_{,\mu} \,,$$

where the comma followed by a Greek index μ denotes partial differentiation with respect to the coordinate x^μ. We also have

$$\dot{x} \cdot \ddot{x} = 0 \quad \text{or} \quad u \cdot \dot{u} = 0 \,,$$

since $u^2 = 1$, and

$$\begin{aligned}
& u_\mu u^\mu{}_{,\nu} = 0 \,, \quad \dot{u}_\mu = u_{\mu,\nu} u^\nu \,, \quad u_\mu u^\mu{}_{,i} = 0 \,, \\
& x^\mu{}_{,i} \xi^i{}_{,\nu} + \dot{x}^\mu \tau_{,\nu} = \delta^\mu{}_\nu \,, \\
& \xi^i{}_{,\mu} x^\mu{}_{,j} = \delta^i_j \,, \quad \xi^i{}_{,\mu} \dot{x}^\mu = 0 \,, \\
& \tau_{,\mu} x^\mu{}_{,i} = 0 \,, \quad \tau_{,\mu} \dot{x}^\mu = 1 \,, \\
& P_{\mu\nu} \dot{x}^\nu{}_{,i} = P_{\mu\nu} u^\nu{}_{,i} = u_{\mu,i} \,.
\end{aligned}$$

We now define the rate of strain tensor for the medium

$$\begin{aligned}
r_{\mu\nu} &:= \dot{\gamma}_{ij} \xi^i{}_{,\mu} \xi^j{}_{,\nu} \\
&= \big(\dot{P}_{\sigma\tau} x^\sigma{}_{,i} x^\tau{}_{,j} + P_{\sigma\tau} \dot{x}^\sigma{}_{,i} x^\tau{}_{,j} \\
&\qquad + P_{\sigma\tau} x^\sigma{}_{,i} \dot{x}^\tau{}_{,j}\big) \, \xi^i{}_{,\mu} \xi^j{}_{,\nu} \\
&= -(\dot{u}_\sigma u_\tau + u_\sigma \dot{u}_\tau)(\delta^\sigma{}_\mu - u^\sigma \tau_{,\mu}) \\
&\qquad \times (\delta^\tau{}_\nu - u^\tau \tau_{,\nu}) \\
&\quad + u_{\tau,i} \xi^i{}_{,\mu} (\delta^\tau{}_\nu - u^\tau \tau_{,\nu}) \\
&\quad + (\delta^\sigma{}_\mu - u^\sigma \tau_{,\mu}) u_{\sigma,j} \xi^j{}_{,\nu} \\
&= -\,(\dot{u}_\mu u_\nu + u_\mu \dot{u}_\nu - \dot{u}_\mu \tau_{,\nu} - \tau_{,\mu} \dot{u}_\nu) \\
&\quad + u_{\nu,\mu} - \dot{u}_\nu \tau_{,\mu} + u_{\mu,\nu} - \dot{u}_\mu \tau_{,\nu} \\
&= -u_{\mu,\sigma} u^\sigma u_\nu - u_\mu u^\sigma u_{\nu,\sigma} + u_{\nu,\mu} + u_{\mu,\nu} \\
&= P_\mu{}^\sigma P_\nu{}^\tau (u_{\sigma,\tau} + u_{\tau,\sigma}) \,.
\end{aligned}$$

This is to be compared with (7.41) to justify calling it the rate of strain tensor. At any event x^μ, it lies entirely in the instantaneous HOS of the particle ξ^i that happens to coincide with that event.

Note in passing that this generalizes to curved spacetimes. We define

$$r_{\mu\nu} := \dot{\gamma}_{ij}\xi^i{}_{,\mu}\xi^j{}_{,\nu} , \tag{7.47}$$

as before, noting that it is a tensor, since γ_{ij}, $\dot{\gamma}_{ij}$, ξ^i, and ξ^j are scalars under change of coordinates. At any x, there are coordinates such that $g_{\mu\nu,\sigma}|_x = 0$, whence covariant derivatives with respect to the Levi-Civita connection are just coordinate derivatives at x, and it follows immediately that

$$r_{\mu\nu} = P_\mu{}^\sigma P_\nu{}^\tau (u_{\sigma;\tau} + u_{\tau;\sigma}) , \tag{7.48}$$

where semicolons denote covariant derivatives and $P^{\mu\nu}$ is given by

$$P^{\mu\nu} = g^{\mu\nu} - \dot{x}^\mu \dot{x}^\nu ,$$

for metric $g^{\mu\nu}$.

So in either special or general relativity, the result

$$\begin{aligned} r_{\mu\nu} &:= \dot{\gamma}_{ij}\xi^i{}_{,\mu}\xi^j{}_{,\nu} \\ &= P_\mu{}^\sigma P_\nu{}^\tau (u_{\sigma;\tau} + u_{\tau;\sigma}) \end{aligned} \tag{7.49}$$

expresses the rate of strain tensor in terms of covariant derivatives of the 4-velocity field of the medium. We now characterise relativistic rigid motion by

$$r_{\mu\nu} = 0 , \quad \dot{\gamma}_{ij} = 0 . \tag{7.50}$$

Once again, we observe that the criterion for rigid motion, namely, $r_{\mu\nu} = 0$, is independent of the coordinates, because $r_{\mu\nu}$ is a tensor, even in a curved spacetime.

7.6 Examples of Rigid Motion

The next problem is to find some examples, restricting to the flat spacetime of special relativity now. We choose an arbitrary particle in the medium and let it be the origin of the labels ξ^i. The problem here is to choose these labels smoothly throughout the medium. Let the worldline $x^\mu(0,\tau)$ of the point $\xi^i = 0$ be arbitrary (but timelike). We introduce a local rest frame for the particle, characterized by an orthonormal triad $n_i{}^\mu(\tau)$

$$\begin{aligned} &n_i \cdot n_j = -\delta_{ij} , \quad n_i \cdot u_0 = 0 , \\ &u_0^2 = 1 , \quad u_0{}^\mu := \dot{x}^\mu(0,\tau) . \end{aligned}$$

We now assume that the worldlines of all the other particles of the medium can be given by

$$x^\mu(\xi,\tau) = x^\mu(0,\sigma) + \xi^i n_i{}^\mu(\sigma) , \tag{7.51}$$

where σ is a certain function of the ξ^i and τ to be determined. On the left, τ is the proper time of the particle labeled by ξ. To achieve a relation of this type, given τ and ξ, we must find the unique proper time σ of the particle $\xi = 0$ such that the point $x^\mu(\xi,\tau)$ is simultaneous with the event $x^\mu(0,\sigma)$ in the instantaneous rest frame of the particle $\xi = 0$. Then the label ξ^i for our particle is defined by the above relation. There is indeed an assumption here, namely that these ξ^i really do label particles. That is, if we look at events with the same ξ^i but varying τ, we are assuming that we do follow a single particle. It is unlikely that all motions of the medium could be expressed like this, but we can obtain some rigid motions, as we shall discover.

To determine the function $\sigma(\xi^i,\tau)$, write

$$u^\mu = \dot{x}^\mu(\xi,\tau) = \left(u_0{}^\mu + \xi^i \dot{n}_i{}^\mu\right)\dot{\sigma} ,$$

all arguments being suppressed in the final expression. Here and in what follows, it is to be understood that dots over u_0 and the n_i denote differentiation with respect to σ, while the dot over σ denotes differentiation with respect to τ.

In order to proceed further, one must expand $\dot{n}_i$ in terms of the orthonormal tetrad u_0, n_i

$$\dot{n}_i{}^\mu = a_{0i}u_0{}^\mu + \Omega_{ij}n_j{}^\mu . \tag{7.52}$$

The coefficients a_{0i} are determined, from the identity

$$\dot{n}_i \cdot u_0 + n_i \cdot \dot{u}_0 = 0 ,$$

to be just the components of the absolute acceleration of the particle $\xi = 0$ in its local rest frame (see an example in (7.20))

$$a_{0i} = -n_i \cdot \dot{u}_0 , \tag{7.53}$$

and the identity

$$\dot{n}_i \cdot n_j + n_i \cdot \dot{n}_j = 0$$

tells us that Ω_{ij} is antisymmetric

$$\Omega_{ij} = -\Omega_{ji} \,.$$

We now have

$$u^\mu = \left[\left(1 + \xi^i a_{0i}\right) u_0{}^\mu + \xi^i \Omega_{ij} n_j{}^\mu\right] \dot{\sigma} \,. \tag{7.54}$$

But

$$1 = u^2 = \left[\left(1 + \xi^i a_{0i}\right)^2 - \xi^i \xi^j \Omega_{ik} \Omega_{jk}\right] \dot{\sigma}^2 \,,$$

whence

$$\dot{\sigma} = \left[\left(1 + \xi^i a_{0i}\right)^2 - \xi^i \xi^j \Omega_{ik} \Omega_{jk}\right]^{-1/2} \,. \tag{7.55}$$

The right-hand side of this equation is a function solely of σ and the ξ^i. Therefore the equation may be integrated along each worldline $\xi = \text{const.}$, subject to the boundary condition

$$\sigma(\xi, 0) = 0 \,.$$

We shall, in particular, have the necessary condition

$$\sigma(0, \tau) = \tau \,.$$

Note that the medium must be confined to regions where

$$\left(1 + \xi^i a_{0i}\right)^2 > \xi^i \Omega_{ik} \xi^j \Omega_{jk} \quad (\geq 0) \,, \tag{7.56}$$

otherwise some of its component particles will be moving faster than light.

We also note that (σ, ξ^i) are semi-Euclidean coordinates for an observer with worldline $x^\mu(0, \sigma)$, moving with the base particle $\xi = 0$. This generalizes the construction of Sect. 7.2 to the case of a general 3-D acceleration. Indeed, the (σ, ξ^i) satisfy the conditions laid down at the beginning of Sect. 7.2.

What we are doing here is to label the particle ξ^i by its spatial coordinates ξ^i in the semi-Euclidean system moving with the particle $\xi = 0$. Geometrically, we have the worldline of the arbitrarily chosen particle O at the origin, namely, $x^\mu(0, \sigma)$, with σ its proper time. We have another worldline $x^\mu(\xi^i, \tau)$ of a particle P labeled by ξ, with proper time τ. For given τ, we seek σ such that $x^\mu(\xi^i, \tau)$ is in the hyperplane of simultaneity of O at its proper time σ. Then (ξ^i) is the position of P in the tetrad moving with O. Indeed, $\{\xi^i\}$ are the space coordinates of P relative to O in that frame.

We can now calculate the proper metric of the medium. We have

$$n_i \cdot u = \Omega_{ij} \xi^j \dot{\sigma} \,, \tag{7.57}$$

$$x^\mu{}_{,i} = n_i{}^\mu + (u_0{}^\mu + \xi^j \dot{n}_j{}^\mu) \sigma_{,i} = n_i{}^\mu + u^\mu \dot{\sigma}^{-1} \sigma_{,i} \,,$$

and hence,

$$u_\mu x^\mu{}_{,i} = \Omega_{ij} \xi^j \dot{\sigma} + \dot{\sigma}^{-1} \sigma_{,i} \,,$$

whereupon we have the following deduction:

$$\begin{aligned}
\gamma_{ij} &= P_{\mu\nu} x^\mu{}_{,i} x^\nu{}_{,j} \\
&= -\delta_{ij} + \Omega_{ik} \xi^k \sigma_{,j} + \Omega_{jk} \xi^k \sigma_{,i} + \dot{\sigma}^{-2} \sigma_{,i} \sigma_{,j} \\
&\quad - \left(\Omega_{ik} \xi^k \dot{\sigma} + \dot{\sigma}^{-1} \sigma_{,i}\right) \left(\Omega_{jl} \xi^l \dot{\sigma} + \dot{\sigma}^{-1} \sigma_{,j}\right) \\
&= -\delta_{ij} - \dot{\sigma}^2 \Omega_{ik} \Omega_{jl} \xi^k \xi^l \\
&= -\delta_{ij} - \frac{\Omega_{ik} \Omega_{jl} \xi^k \xi^l}{(1 + \xi^m a_{0m})^2 - \xi^n \xi^r \Omega_{ns} \Omega_{rs}} \,,
\end{aligned} \tag{7.58}$$

using the above expression (7.55) for $\dot{\sigma}$.

From this expression we see that there are two ways in which the motion of the medium can be rigid:

- All the Ω_{ij} are zero.
- All the Ω_{ij} and all the a_{0i} are constants, independent of σ.

In the second case the motion is one of a six-parameter family, with the Ω_{ij} and the a_{0i} as parameters. These special motions are sometimes called superhelical motions. One example, constant rotation about a fixed axis, is discussed in Sect. 7.8, while all superhelical motions will be characterized in Sect. 7.9. But first we consider the case where all the Ω_{ij} are zero.

7.7 Rigid Motion Without Rotation

Saying that the Ω_{ij} are all zero amounts to saying that the triad $n_i^{\ \mu}$ is Fermi–Walker (FW) transported along the worldline of the particle $\xi = 0$. Let us see briefly what this means.

If $u_0(\sigma)$ is the 4-velocity of the worldline, the equation for FW transport of a contravector A^μ along the worldline is

$$\dot{A} = -(A \cdot \dot{u}_0)u_0 + (A \cdot u_0)\dot{u}_0 \,. \tag{7.59}$$

This preserves inner products, i. e., if A and B are FW transported along the worldline, then $A \cdot B$ is constant along the worldline. Furthermore, the tangent vector u_0 to the worldline is itself FW transported along the worldline, and if the worldline is a spacetime geodesic (a straight line in Minkowski coordinates), then FW transport is the same as parallel transport.

Now recall that the Ω_{ij} were defined by

$$\dot{n}_i^{\ \mu} = a_{0i}u_0^{\ \mu} + \Omega_{ij}n_j^{\ \mu} \,. \tag{7.60}$$

When $\Omega_{ij} = 0$, this becomes

$$\dot{n}_i^{\ \mu} = a_{0i}u_0^{\ \mu} \,. \tag{7.61}$$

This is indeed the FW transport equation for $n_i^{\ \mu}$, found by inserting $A = n_i$ into (7.59), because we insist on $n_i \cdot u_0 = 0$ and we have $a_{0i} = -n_i \cdot \dot{u}_0$ (7.53).

In fact, the orientation in spacetime of the local rest frame triad $n_i^{\ \mu}$ cannot be kept constant along a worldline unless that worldline is straight (we are referring to flat spacetimes here). Under FW transport, however, the triad remains as constantly oriented, or as rotationless, as possible, in the following sense: at each instant of time σ, the triad is subjected to a pure Lorentz boost without rotation in the instantaneous hyperplane of simultaneity. (On a closed orbit, this process can still lead to spatial rotation of axes upon return to the same space coordinates, an effect known as Thomas precession.) For a general non-Fermi–Walker transported triad, the Ω_{ij} are the components of the angular velocity tensor that describes the instantaneous rate of rotation of the triad in the instantaneous HOS.

Of course, given any triad $n_i^{\ \mu}$ at one point on the worldline, it is always possible to FW transport it to other points by solving (7.59). We are then saying that motions that can be given by (7.51), namely,

$$x^\mu(\xi, \tau) = x^\mu(0, \sigma) + \xi^i n_i^{\ \mu}(\sigma) \,, \tag{7.62}$$

where the ξ^i are assumed to label material particles in the medium, are rigid in the sense of the criterion given above. Furthermore, the proper geometry of the medium given by the proper metric γ_{ij} in (7.39) is then flat, i. e.,

$$\gamma_{ij} = -\delta_{ij} \,.$$

As attested by (7.57), we also have

$$n_i \cdot u = 0 \,, \tag{7.63}$$

so that the instantaneous HOS of the particle at $\xi = 0$ is an instantaneous HOS for all the other particles of the medium as well, and the triad $n_i^{\ \mu}$ serves to define a rotationless rest frame for the whole medium. In other words, the coordinate system defined by the particle labels ξ^i may itself be regarded as being FW transported, and all the particles of the medium have a common designator of simultaneity in the parameter σ. In the semi-Euclidean system, σ is taken to be the time coordinate.

Put another way, (7.63) says that the $n_i(\sigma)$ are in fact orthogonal to the worldline of the particle labeled by ξ^i at the value of τ corresponding to σ. This happens because $u(\xi, \tau) = u_0(0, \sigma)$. In words, the 4-velocity of particle ξ at its proper time τ is the same as the 4-velocity of the base particle when it is simultaneous with the latter in the reckoning of the base particle (quite a remarkable thing).

Because σ is not generally equal to τ, however, it is not possible for the particles to have a common synchronization of standard clocks. The relation between σ and τ is given by (7.55) as

$$\dot{\sigma} = \left(1 + \xi^i a_{0i}\right)^{-1} \,.$$

We can thus find the absolute acceleration a_i of an arbitrary particle in terms of a_{0i} and the ξ^i

$$\begin{aligned} a_i &= -n_i \cdot \dot{u} = -n_i \cdot \frac{\partial u}{\partial \sigma}\dot{\sigma} \\ &= -\dot{\sigma} n_i \cdot \frac{\partial}{\partial \sigma}\left[\left(1 + \xi^j a_{0j}\right) u_0 \dot{\sigma}\right] \\ &= -\dot{\sigma} n_i \cdot \dot{u}_0 \\ &= \frac{a_{0i}}{1 + \xi^j a_{0j}} \,. \end{aligned} \tag{7.64}$$

Here we have used the fact that $u = \left(1 + \xi^j a_{0j}\right) u_0 \dot{\sigma} = u_0$. We see that, although the motion is rigid and rotationless in the sense described above, not all parts of the medium are subject to the same acceleration.

It is important to note that, when we find ξ^i and σ, they constitute semi-Euclidean coordinates (adapted to $\xi = 0$) for the point $x^\mu(\xi, \tau)$ whether or not that point follows a particle for fixed ξ. What we have here are material particles that follow all these points with fixed ξ, for a whole 3-D range of values of ξ.

In these coordinates, the metric tensor takes the form

$$g_{00} = \left.\frac{\partial x^\mu}{\partial \sigma}\right|_\xi \left.\frac{\partial x^\nu}{\partial \sigma}\right|_\xi \eta_{\mu\nu} = u^2 \dot{\sigma}^{-2} = \left(1 + \xi^i a_{0i}\right)^2 ,$$

$$g_{i0} = g_{0i} = \left.\frac{\partial x^\mu}{\partial \xi^i}\right|_\sigma \left.\frac{\partial x^\nu}{\partial \sigma}\right|_\xi \eta_{\mu\nu} = (n_i \cdot u)\dot{\sigma}^{-1} = 0 ,$$

$$g_{ij} = \left.\frac{\partial x^\mu}{\partial \xi^i}\right|_\sigma \left.\frac{\partial x^\nu}{\partial \xi^j}\right|_\sigma \eta_{\mu\nu} = n_i \cdot n_j = -\delta_{ij} ,$$

which has a simple diagonal structure. We note that this metric becomes static, i. e., time-independent, with the parameter σ playing the role of time, in the special case in which the absolute acceleration of each particle is constant. This should be compared with (7.16) and (7.17), and also (7.21).

We conclude that this rigid motion possesses only the three degrees of freedom that the particle $\xi = 0$ itself possesses. The base particle $\xi = 0$ can move any way it wants, but the rest of the medium must then follow in a well defined way.

7.8 Rigid Rotation

The simplest example of a medium undergoing rigid rotation is obtained by choosing

$$a_{0i} = 0 , \quad \Omega_{12} = \omega , \quad \Omega_{23} = 0 = \Omega_{31} .$$

The worldline of the particle at $\xi = 0$ is then straight, but the worldlines of all the other particles are helices of constant pitch. We have

$$\dot{\sigma} = \left\{1 - \omega^2 \left[(\xi^1)^2 + (\xi^2)^2\right]\right\}^{-1/2}$$

and the proper metric of the medium takes the form

$$(\gamma_{ij}) = \begin{pmatrix} -1 - (\dot{\sigma}\omega\xi^2)^2 & +(\dot{\sigma}\omega)^2\xi^1\xi^2 & 0 \\ +(\dot{\sigma}\omega)^2\xi^1\xi^2 & -1 - (\dot{\sigma}\omega\xi^1)^2 & 0 \\ 0 & 0 & -1 \end{pmatrix} .$$

Relabeling the particles by means of three new coordinates r, θ, z given by

$$\xi^1 = r\cos\theta , \quad \xi^2 = r\sin\theta , \quad \xi^3 = z , \qquad (7.65)$$

the proper metric of the rotating medium takes the form

$$-\operatorname{diag}\left(1, \frac{r^2}{1 - \omega^2 r^2}, 1\right) .$$

Indeed, we have

$$\dot{\sigma}^2 = \frac{1}{1 - \omega^2 r^2} ,$$

whence

$$\begin{aligned} -\gamma_{r'r'} &= -\frac{\partial \xi^i}{\partial r}\frac{\partial \xi^j}{\partial r}\gamma_{ij} \\ &= \cos^2\theta\left[1 + (\dot{\sigma}\omega r)^2 \sin^2\theta\right] \\ &\quad - 2(\dot{\sigma}\omega r)^2 \sin^2\theta\cos^2\theta \\ &\quad + \sin^2\theta\left[1 + (\dot{\sigma}\omega r)^2\cos^2\theta\right] \\ &= 1 , \end{aligned}$$

$$\begin{aligned} -\gamma_{r\theta} = -\gamma_{\theta r'} &= -\frac{\partial \xi^i}{\partial r}\frac{\partial \xi^j}{\partial \theta}\gamma_{ij} \\ &= -r\sin\theta\cos\theta\left[1 + (\dot{\sigma}\omega r)^2\sin^2\theta\right] \\ &\quad - r(\dot{\sigma}\omega r)^2 \sin\theta\cos^3\theta \\ &\quad + r(\dot{\sigma}\omega r)^2\sin^3\theta\cos\theta \\ &\quad + r\sin\theta\cos\theta\left[1 + (\dot{\sigma}\omega r)^2\cos^2\theta\right] \\ &= 0 , \end{aligned}$$

$$\gamma_{rz} = \gamma_{zr'} = \frac{\partial \xi^i}{\partial r}\frac{\partial \xi^j}{\partial z}\gamma_{ij} = 0 ,$$

$$\begin{aligned} -\gamma_{\theta\theta} &= -\frac{\partial \xi^i}{\partial \theta}\frac{\partial \xi^j}{\partial \theta}\gamma_{ij} \\ &= r^2\sin^2\theta\left[1 + (\dot{\sigma}\omega r)^2\sin^2\theta\right] \\ &\quad + 2r^2(\dot{\sigma}\omega r)^2\sin^2\theta\cos^2\theta \\ &\quad + r^2\cos^2\theta\left[1 + (\dot{\sigma}\omega r)^2\cos^2\theta\right] \\ &= r^2\left[1 + (\dot{\sigma}\omega r)^2\right] \\ &= r^2\left(1 + \frac{\omega^2 r^2}{1 - \omega^2 r^2}\right) = \frac{r^2}{1 - \omega^2 r^2} , \end{aligned}$$

$$\gamma_{\theta z} = \gamma_{z\theta} = \frac{\partial \xi^i}{\partial \theta}\frac{\partial \xi^j}{\partial z}\gamma_{ij} = 0\,,$$
$$\gamma_{zz} = \frac{\partial \xi^i}{\partial z}\frac{\partial \xi^j}{\partial z}\gamma_{ij} = -1\,.$$

In terms of these coordinates the proper distance δs between two particles separated by displacements δr, $\delta\theta$, and δz therefore takes the form

$$\delta s^2 = (\delta r)^2 + \frac{r^2}{1-\omega^2 r^2}(\delta\theta)^2 + (\delta z)^2\,.$$

We are merely applying (7.38) for the new particle labels. This gives the distance of one particle as reckoned in the instantaneous rest frame of the neighboring particle. The second term on the right of this equation may be understood as arising from relativistic contraction.

At first sight, it may look odd to find that, when a disc of radius r is set spinning with angular frequency ω about its axis, so that radial distances are unaffected by relativistic contraction, distances in the direction of rotation contract in such a way that the circumference of the disc gets reduced to the value $2\pi R\sqrt{1-\omega^2R^2}$. It appears to contradict the Euclidean nature of the ordinary 3-space that the disc inhabits!

This can be clarified as follows. Suppose A and B are two neighboring particles at distance R from the center and with labels θ and $\theta+\delta\theta$. When the disk is not rotating, the proper distance between them as reckoned by either in its ICIF is $R\delta\theta$. When the disk is rotating, the expression for γ_{ij} tells us that the proper distance between them in the new ICIF will increase to $R\delta\theta/(1-\omega^2R^2)^{1/2}$. Seen by an inertial observer moving with the center of the disk, this separation will thus be $R\delta\theta$, as before, and there will be no contradiction with the edicts of Euclidean geometry. This shows that the matter between A and B is stretched in the sense of occupying a greater proper distance as judged in an ICIF moving with either A or B.

The above discussion does assume that θ labels the material particles! And this follows from the relations in (7.65) and the fact that ξ^1, ξ^2, ξ^3 label the particles. It would be easy to miss this point. There remains therefore the question as to whether any association of material particles could have, or is likely to have this motion.

We note that the medium must be confined to regions where $r<\omega^{-1}$ and that its motion will not be rigid if ω varies with time. There are no degrees of freedom in this kind of (superhelical) motion: once the medium gets into superhelical motion, it must remain frozen into it if it wants to stay rigid. We also note that the proper geometry of the medium is not flat, i. e., $\gamma_{ij} \neq -\delta_{ij}$.

7.9 Generalized Uniform Acceleration and Superhelical Motions

The notion of uniform acceleration, previously limited to straight line motion, has recently been elegantly generalized by *Friedman* and *Scarr* in a way that allows us to identify all superhelical motions [7.5]. Let us briefly discuss the key points in the context of what has come before.

It will be useful first to review the big picture. If we consider an arbitrary timelike worldline in special relativity, we have seen that we can always find coordinate systems $\{\xi^i, \tau\}$ adapted to that worldline in the following sense:

- The worldline is given by $\xi^i = 0$, $i = 1,2,3$.
- The coordinate τ is equal to the proper time σ along the worldline.
- It is easily shown that the metric is given at any event (ξ^i, τ) by

$$\boxed{\begin{aligned} g_{00} &= \left(1+\xi^i a_{0i}\right)^2 - \xi^i\xi^j\Omega_{ik}\Omega_{jk}\,,\\ g_{0i} &= \xi^j\Omega_{ji} = g_{i0}\,, \quad g_{ij} = -\delta_{ij}\,, \end{aligned}} \tag{7.66}$$

where $\Omega_{ij}(\sigma)$ is an antisymmetric 3×3 matrix describing the rotation of the spatial coordinate axes as one moves along the worldline (see (7.52)) and $a_{0i}(\sigma)$ are the three nonzero components of the acceleration 4-vector in the ICIF.

- The metric reduces to the standard form $\eta_{\mu\nu}$ of the Minkowski metric on the worldline and induces the Euclidean metric on the spacelike hypersurfaces of simultaneity $\tau =$ constant for these coordinates.
- One can also easily show that the connection is given on the worldline itself by

$$\boxed{\Gamma^i_{00} = \Gamma^0_{0i} = \Gamma^0_{i0} = a_{0i}\,, \quad i = 1,2,3\,,} \tag{7.67}$$

$$\boxed{\Gamma^\mu_{ij} = 0\,, \quad \mu = 0,1,2,3, \quad i = 1,2,3\,,} \tag{7.68}$$

$$\boxed{\Gamma^i_{0j} = \Gamma^i_{j0} = \Omega_{ij}\,, \quad i,j = 1,2,3\,,} \tag{7.69}$$

and hence encodes the acceleration of the worldline and the rotation of the spatial coordinate axes.

Choosing spatial coordinate axes such that $\Omega_{ij}(\sigma) = 0$ for all proper times σ along the worldline amounts to selecting a spacelike triad orthogonal to the unit tangent to the worldline, which is just its 4-velocity, at some point, and then FW transporting that triad along the worldline to specify the spatial coordinate axes at other points along the worldline.

This also achieves rigidity of the coordinate system in the following sense. Any two observers sitting at two fixed neighboring space coordinates ξ and $\xi + \delta\xi$ are always the same proper distance apart as measured by either in an ICIF. This in turn means that a ruler satisfying the ruler hypothesis would always correctly indicate spatial coordinate separations when made to sit at fixed spatial coordinates in this system.

In the case of nonrotating spatial coordinate axes, the metric has the form

$$(g_{\mu\tau}) = \begin{pmatrix} (1+\xi^j a_{0j})^2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix},$$

which is static if and only if the absolute acceleration components a_{0i} are constants (independent of proper time along the worldline), a motion generally known as (translational) uniform acceleration.

Let us now see how we can generalize the notion of uniform acceleration and obtain a rigid coordinate system without the need to FW transport the initial ICIF along the worldline.

7.9.1 Definition

Here we closely follow the discussion by *Friedman* and *Scarr* [7.5]. We work in an inertial (laboratory) frame denoted by K. For any timelike worldline, we take the 4-velocity to be the dimensionless unit 4-vector

$$\boldsymbol{u} = (u^0, u^1, u^2, u^3) := \frac{1}{c}\frac{\mathrm{d}x^\mu}{\mathrm{d}\tau},$$

where τ is the proper time, and hence define the 4-acceleration to be

$$a^\mu := c\frac{\mathrm{d}u^\mu}{\mathrm{d}\tau},$$

which has units of acceleration.

We define a uniformly accelerating worldline to be one that satisfies

$$\boxed{c\frac{\mathrm{d}u^\mu}{\mathrm{d}\tau} = A^\mu{}_\nu u^\nu}, \tag{7.70}$$

with some specified initial value $u(0) = u_0$, where $A^\mu{}_\nu$ is a tensor under Lorentz transformations and independent of τ. We also require that the type (2,0) form $\bar{A}$ of this tensor, with components $A_{\mu\nu} := \eta_{\mu\sigma}A^\sigma{}_\nu$, should be antisymmetric, for the following reason. Since $u^2 = 1$ is constant, we must have $a\cdot u = 0$, whence we require

$$0 = \eta_{\mu\nu}a^\mu u^\nu = \eta_{\mu\nu}A^\mu{}_\sigma u^\sigma u^\nu = u^\nu A_{\nu\sigma}u^\sigma,$$

a sufficient condition for which is that the type (2,0) tensor $A_{\mu\nu}$ should be antisymmetric, i. e.,

$$A_{\mu\nu} = -A_{\nu\mu}. \tag{7.71}$$

Equation (7.70) has a unique solution

$$\boxed{u(\tau) = \exp\left(\frac{A\tau}{c}\right)u_0 = \left(\sum_{n=0}^{\infty}\frac{A^n}{n!c^n}\tau^n\right)u_0}, \tag{7.72}$$

where A is type (1,1) tensor. A key motivation for the above definition is that this kind of motion is covariant in the sense that uniformly accelerated motion in one inertial frame is uniformly accelerated motion in every inertial frame. This in turn follows straight from the definition because proper time is an invariant, u is a 4-vector, $A^\mu{}_\nu$ is a tensor (see below), and $A_{\mu\nu}$ will be antisymmetric in every inertial frame if it is so in one.

7.9.2 Tensorial Nature of A and $\bar{A}$

The equation of motion (7.70) has been expressed relative to some arbitrarily chosen inertial frame K. But how would we transform this equation of motion in order to describe the worldline relative to a new inertial frame?

The answer is that we will get *the same equation* expressed relative to the new frame if we transform the object so suggestively written as $A^\mu{}_\nu$ as a type (1,1) tensor. Indeed, if we are to obtain the same equation expressed relative to the new frame, *it has to transform*

like this because the left-hand side of (7.70) transforms as a contravector, and so does u.

Relative to any other choice of laboratory inertial frame K_1 related to K by a homogeneous Lorentz transformation, the acceleration matrix will have the form

$$A_1 = L^{-1}AL\,, \tag{7.73}$$

where L is the homogeneous Lorentz transformation from K to K_1.

Naturally then, the object $\bar{A}$ with components $A_{\mu\nu} := \eta_{\mu\sigma}A^{\sigma}{}_{\nu}$ must transform as a type (2,0) tensor when we rewrite the equation of motion relative to some other inertial frame. Relative to any other choice of laboratory inertial frame K_1 related to K by a homogeneous Lorentz transformation L, type (2,0) acceleration matrix $\bar{A}$ will have the form

$$\bar{A}_1 = L^{\mathrm{T}}\bar{A}L\,. \tag{7.74}$$

Note that $\bar{A}$ is antisymmetric if and only if $L^{\mathrm{T}}\bar{A}L$ is antisymmetric.

7.9.3 Nature of Generalization

Let us see how the above extends the usual definition of uniform acceleration. The first thing is to write down the most general possible matrix versions of $A_{\mu\nu}$ and $A^{\mu}{}_{\nu}$ in the chosen laboratory inertial frame K, bearing in mind the antisymmetry of the former

$$\begin{aligned} A_{\mu\nu}(\boldsymbol{g},\boldsymbol{\omega}) &= \begin{pmatrix} 0 & \boldsymbol{g}^{\mathrm{T}} \\ -\boldsymbol{g} & -c\pi(\boldsymbol{\omega}) \end{pmatrix}\,, \\ A^{\mu}{}_{\nu}(\boldsymbol{g},\boldsymbol{\omega}) &= \begin{pmatrix} 0 & \boldsymbol{g}^{\mathrm{T}} \\ \boldsymbol{g} & c\pi(\boldsymbol{\omega}) \end{pmatrix}\,. \end{aligned} \tag{7.75}$$

Here we have used the notation introduced in [7.5]: $\boldsymbol{g}$ is a 3-component object with units of acceleration and transpose $\boldsymbol{g}^{\mathrm{T}}$, and $\boldsymbol{\omega} = (\omega^1, \omega^2, \omega^3)$ is another 3-component object but this time with units of 1/time, and

$$\pi(\boldsymbol{\omega}) := \varepsilon_{ijk}\omega^k\,, \tag{7.76}$$

with ε_{ijk} the completely antisymmetric Levi-Civita symbol. The factor of c with $\pi(\boldsymbol{\omega})$ just ensures that this entry has units of acceleration. Since $A_{\mu\nu}$ is independent of τ, the same goes for $\boldsymbol{g}$ and $\boldsymbol{\omega}$.

Now when $\boldsymbol{\omega} = 0$, the above definition of uniform acceleration reduces to the usual definition of uniform acceleration in a straight line. Indeed, we then have

$$c\frac{\mathrm{d}u^{\mu}}{\mathrm{d}\tau} = A^{\mu}{}_{\nu}u^{\nu} = \begin{pmatrix} 0 & \boldsymbol{g}^{\mathrm{T}} \\ \boldsymbol{g} & 0 \end{pmatrix}\begin{pmatrix} u^0 \\ \frac{\boldsymbol{u}}{c} \end{pmatrix}\,, \tag{7.77}$$

so that

$$c\frac{\mathrm{d}u^{\mu}}{\mathrm{d}\tau} = (c^{-1}\boldsymbol{g}\cdot\boldsymbol{u}, \boldsymbol{g}u^0)\,, \tag{7.78}$$

since we are taking

$$(u^{\mu}) = (u^0, \boldsymbol{u}/c)\,, \quad u^0 = \gamma(v)\,, \quad \boldsymbol{u} = \gamma(v)\boldsymbol{v}\,.$$

We also note that

$$\frac{\mathrm{d}t}{\mathrm{d}\tau} = \gamma\,,$$

whence

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} = \frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}\tau}\frac{\mathrm{d}\tau}{\mathrm{d}t} = \boldsymbol{g}u^0\gamma^{-1} = \boldsymbol{g}\,. \tag{7.79}$$

Since $\boldsymbol{g}$ is independent of time, this is indeed the usual definition for translational uniform acceleration (TUA). It has solution

$$\boldsymbol{u} = \boldsymbol{g}t + \boldsymbol{u}_0\,, \tag{7.80}$$

where $\boldsymbol{u}_0$ is the value of $\boldsymbol{u}$ at time $t = 0$. It is not difficult to see how this accords with the earlier definition.

Note that the definition of purely TUA does not give rise to a covariant notion of uniform acceleration, since it depends on having $\boldsymbol{\omega} = 0$. This standard notion would thus only be covariant under transformations that preserve this condition, namely, Lorentz boosts in the direction of $\boldsymbol{g}$ and space rotations about the direction of $\boldsymbol{g}$. Rather than being the whole homogeneous Lorentz group, as for the new definition of uniform acceleration, the covariance group would be the little group fixing the space axis along $\boldsymbol{g}$. In fact, the generalized form of uniform acceleration is even covariant under the transformations of the inhomogeneous Lorentz group (Poincaré group), including spacetime translations.

The type (1,1) tensor A and the associated antisymmetric type (2,0) tensor $\bar{A}$ are both referred to as the acceleration tensor. A uniformly accelerated motion is uniquely defined by its acceleration tensor A and its initial 4-velocity u_0. The associated worldline $\hat{x}(\tau)$ can be found if we know the initial position $\hat{x}(0)$. The basic equation (7.70) along with an initial value $u(0) = u_0$ can be solved exactly to obtain $u(\tau)$, and the resulting expression is easily integrated to obtain $\hat{x}(\tau)$ if we have the initial value $\hat{x}(0)$. The details can be found in [7.5].

7.9.4 Coordinate Frame for Generalized Uniform Acceleration

Once again, we follow *Friedman* and *Scarr* for this construction [7.5]. The aim will be to construct a coordinate frame K' such that the observer with uniform acceleration and proper time τ has worldline $(c\tau, 0, 0, 0)$. We shall do this in a mathematically natural way and we shall find that the result is a rigid frame in the sense that two nearby particles sitting at fixed space coordinates in K' are always the same distance apart as measured in the instantaneously comoving inertial frame of either.

Note that we already have a rigid frame, obtained by FW transporting a space triad along the observer worldline and using the construction described in detail earlier. In the present case, we shall obtain a generally different rigid coordinate frame that depends crucially on the observer worldline being generated as in (7.70) by a constant acceleration tensor. We shall once again transport a space triad along the observer worldline, but the transport will be a generalization of FW transport, although it will be a generalization *only* for the case of generalized uniform acceleration (GUA), and only reduce to it in the special case where we have the purely translational form of uniform acceleration in some inertial frame in which the observer comes to rest at some event.

Consider the worldline $\hat{x}(\tau)$ of a uniformly accelerating observer in the sense of (7.70), with motion determined by a constant acceleration matrix $\tilde{A} = (A_{\mu\nu})$, initial 4-velocity $u(0)$, and initial position $\hat{x}(0)$. The first step is to define an ICIF K_τ at each proper time τ, specified by a tetrad $\lambda(\tau) = \{\lambda_{(\kappa)}(\tau)\}_{\kappa=0,1,2,3}$. To do so we choose an initial ICIF K_0 with origin at $\hat{x}(0)$, specified by a tetrad $\hat{\lambda} = \{\hat{\lambda}_{(\kappa)}\}_{\kappa=0,1,2,3}$, where as usual we take $\hat{\lambda}_{(0)} = u(0)$ and $\{\hat{\lambda}_{(i)}\}_{i=1,2,3}$ can be chosen arbitrarily to complete the tetrad.

We must now transport this initial tetrad along the worldline. Previously we used FW transport and obtained a perfectly good rigid coordinate frame in that way. However, there is a mathematically more natural way to transport our space triad in the present case. For $\tau > 0$, we define K_τ by requiring the origin of K_τ at time τ to be $\hat{x}(\tau)$ and requiring the basis of K_τ to be the unique solution $\lambda(\tau) = \{\lambda_{(\kappa)}(\tau)\}_{\kappa=0,1,2,3}$ of the initial value problem

$$\boxed{c\frac{\mathrm{d}\lambda^{\mu}_{(\kappa)}}{\mathrm{d}\tau} = A^{\mu}{}_{\nu}\lambda^{\nu}_{(\kappa)}\,, \quad \lambda_{(\kappa)}(0) = \hat{\lambda}_{(\kappa)}}\,. \tag{7.81}$$

Since $\hat{\lambda}_{(0)} = u(0)$ and $u(\tau)$ satisfies this differential equation according to (7.70), we deduce that $\lambda_{(0)}(\tau) = u(\tau)$ for all values of τ. Furthermore, the type (1,1) tensor A is a constant matrix, and we can immediately solve the system (7.81) to obtain

$$\lambda(\tau) = \exp\left(\frac{A\tau}{c}\right)\hat{\lambda}\,. \tag{7.82}$$

Note also that this kind of transport is an isometry, that is, it preserves the Lorentzian scalar product. To see this, suppose that v and w are any two 4-vectors at $\hat{x}(0)$ and solve (7.81) to obtain vector fields $v(\tau)$ and $w(\tau)$ along $\hat{x}(\tau)$. Then consider

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}[v(\tau)\cdot w(\tau)] &= [Av(\tau)]\cdot[w(\tau)] \\ &\quad + [v(\tau)]\cdot[Aw(\tau)] \\ &= \eta_{\mu\sigma}A^{\mu}{}_{\nu}v^{\nu}w^{\sigma} + \eta_{\mu\sigma}v^{\mu}A^{\sigma}{}_{\nu}w^{\nu} \\ &= A_{\sigma\nu}v^{\nu}w^{\sigma} + A_{\mu\nu}v^{\mu}w^{\nu} \\ &= A_{\mu\nu}(v^{\nu}w^{\mu} + v^{\mu}w^{\nu}) = 0\,,\end{aligned} \tag{7.83}$$

due to the antisymmetry of the type (2,0) tensor $\tilde{A}$. Interestingly, we do not use the constancy of the matrix A in this proof, only the differential relations that v and w must satisfy, so this kind of transport is isometric for quite general, possibly time-varying acceleration matrices A, provided that the associated matrix $\tilde{A}$ is antisymmetric.

The fact that this transport is isometric is important, because it means that the solution to (7.81) will be orthonormal right along the observer worldline, i. e., it will be a tetrad. We shall examine the resulting coordinate system in a moment. Before doing so, it is interesting to rewrite (7.81) in a slightly different way. To begin with, let us think of our initial frame $\hat{\lambda}$ as a matrix with columns

$$\hat{\lambda} = \left(\hat{\lambda}_{(0)}\hat{\lambda}_{(1)}\hat{\lambda}_{(2)}\hat{\lambda}_{(3)}\right)\,, \tag{7.84}$$

where each column comprises the components $\hat{\lambda}^{\mu}_{(\kappa)}$ of the given tetrad 4-vector expressed relative to the inertial (laboratory) frame K. This matrix maps the basis

$$e_0 := \begin{pmatrix}1\\0\\0\\0\end{pmatrix}, \quad e_1 := \begin{pmatrix}0\\1\\0\\0\end{pmatrix}, \quad e_2 := \begin{pmatrix}0\\0\\1\\0\end{pmatrix}, \quad e_3 := \begin{pmatrix}0\\0\\0\\1\end{pmatrix}, \tag{7.85}$$

of K to the basis we have chosen for the initial ICIF K_0. Now define the new matrix

$$\tilde{A} := \hat{\lambda}^{-1} A \hat{\lambda} \,, \tag{7.86}$$

the representation of the type (1,1) tensor A relative to the basis of the initial ICIF K_0. Put another way, we now have

$$\hat{\lambda}\tilde{A} = A\hat{\lambda} \,, \tag{7.87}$$

or in component form

$$\hat{\lambda}^{\nu}_{(\gamma)} \tilde{A}^{(\gamma)}{}_{(\kappa)} = A^{\nu}{}_{\sigma} \hat{\lambda}^{\sigma}_{(\kappa)} \,. \tag{7.88}$$

The point is that we can now write

$$c\frac{\mathrm{d}\lambda^{\mu}_{(\kappa)}}{\mathrm{d}\tau} = \left[\exp\left(\frac{A\tau}{c}\right)\right]^{\mu}{}_{\nu} A^{\nu}{}_{\sigma} \hat{\lambda}^{\sigma}_{(\kappa)}$$
$$= \left[\exp\left(\frac{A\tau}{c}\right)\right]^{\mu}{}_{\nu} \hat{\lambda}^{\nu}_{(\gamma)} \tilde{A}^{(\gamma)}{}_{(\kappa)} \,,$$

whence

$$\boxed{c\frac{\mathrm{d}\lambda^{\mu}_{(\kappa)}}{\mathrm{d}\tau} = \lambda^{\mu}_{(\gamma)} \tilde{A}^{(\gamma)}{}_{(\kappa)}} \,. \tag{7.89}$$

This is just a slightly different, but equivalent version of (7.81). It is useful for drawing the parallel with previous constructions, as we shall see, and also for showing that this kind of transport generalizes FW transport in the case of purely TUA in the initial ICIF $K_0 = \hat{\lambda}$.

These considerations generalize easily to the case of acceleration matrices that are not constant along the observer worldline. It can be shown that we can always find a matrix $A^{\mu}{}_{\nu}$ such

$$c\frac{\mathrm{d}\lambda^{\mu}{}_{(0)}}{\mathrm{d}\tau} = A^{\mu}{}_{\nu} \lambda^{\nu}{}_{(0)} \,, \tag{7.90}$$

although it will not generally be constant, with the further property that

$$c\frac{\mathrm{d}\lambda^{\mu}{}_{(i)}}{\mathrm{d}\tau} = A^{\mu}{}_{\nu} \lambda^{\nu}{}_{(i)} \,, \quad i = 1,2,3 \,, \tag{7.91}$$

whatever smooth choice of space tetrad $\{\lambda^{\mu}{}_{(i)}\}_{i=1,2,3}$ has been made along the worldline (but noting that the matrix A then depends on that choice).

Let us now see how to set up coordinates $\{y^{(\mu)}\}$ adapted to the generalized uniformly accelerating observer worldline. For any event X with coordinates x^{μ} in K, we find a proper time τ for the observer for which $\hat{x}(\tau)$ is simultaneous with X in the ICIF K_{τ}. We then define the zero (or time) coordinate of X in the proposed accelerating frame K' to be $y^{(0)} = c\tau$. Note that all events in this particular HOS for K_{τ} will be attributed the same time coordinate $c\tau$.

Put another way, K' is borrowing the hyperplanes of simultaneity of an instantaneously comoving inertial observer, so given that we obtain the tetrad field along the observer worldline by the isometric propagation (7.81) (or (7.89)), we have a standard construction of semi-Euclidean coordinates as described previously. And as in our earlier constructions, we still have the problem that such hyperplanes can intersect off the observer worldline. These coordinates will generally only be valid on some neighborhood of the worldline, and not throughout the whole of spacetime.

Now the displacement $\bar{y}$ of X relative to the observer at proper time τ can be expressed in terms of the space triad $\{\lambda_{(i)}(\tau)\}_{i=1,2,3}$, since

$$\bar{y} \cdot u(\tau) = [x - \hat{x}(\tau)] \cdot u(\tau) = 0 \,,$$

by the specific choice of τ. Hence, there are $y^{(i)} \in \mathbb{R}$, $i = 1,2,3$, such that

$$\bar{y} = y^{(i)} \lambda_{(i)}(\tau) \,.$$

In short, we have found τ such that

$$\boxed{x^{\mu} = \hat{x}^{\mu}(\tau) + y^{(i)} \lambda_{(i)}(\tau)} \,. \tag{7.92}$$

The coordinates of event X in the coordinate frame K' will then be defined as $(c\tau, y^{(1)}, y^{(2)}, y^{(3)})$, and the relation (7.92) tells us how to convert from these coordinates to the original laboratory coordinate system K.

7.9.5 Rigidity

We shall show that this is a rigid coordinate system in the sense of Sect. 7.4. It is important to understand that this notion of rigidity is *not* the same as saying that the geometry of the hyperplanes of simultaneity is Euclidean, which is true by construction. Imagine two particles A and B at rest relative to the space coordinates of the proposed accelerating frame K', with worldlines of the form $\{(c\tau, \boldsymbol{y}_A) : \tau \in \mathbb{R}\}$ and $\{(c\tau, \boldsymbol{y}_B) : \tau \in \mathbb{R}\}$, respectively. At a given τ, the particles lie in the same

hyperplane of simultaneity of the inertial frame K_τ, and since this is also the HOS adopted in K', the proper distance between A and B at coordinate time τ in the K' system is just the length of the vector $[y_B^{(i)} - y_A^{(i)}]\lambda_{(i)}(\tau_0)$, namely,

$$\sqrt{\delta_{ij}\left[y_B^{(i)} - y_A^{(i)}\right]\left[y_B^{(j)} - y_A^{(j)}\right]}\,.$$

The fact that this is independent of τ does not prove rigidity.

Rigidity according to Sect. 7.4 means that neighboring worldlines with fixed space coordinates are always the same proper distance apart *as measured in the instantaneous rest frame of either*. We need therefore to examine the proper distance between A and B as measured in the instantaneous rest frame of A, for example, which depends on the motion of A. The frame K' is nevertheless rigid in this sense, ultimately because the acceleration matrix is constant, but this is not obvious and requires more work.

We have already done this work in Sect. 7.6, however. The key will be (7.89). Recall that A is a constant matrix, i. e., independent of proper time τ, and so of course is the matrix $\hat{\lambda}$ of (7.84). This means that the matrix $\tilde{A}$ in (7.89) is also independent of τ. But (7.89) corresponds exactly to the key relation (7.52) in Sect. 7.6, namely,

$$c\dot{n}_i^{\ \mu} = a_{0i}u^\mu + c\Omega_{ij}n_j^{\ \mu}\,, \tag{7.93}$$

where we have reinstated c and replaced the notation u_0 for the 4-velocity of the observer by the present notation u, recalling that we made the latter dimensionless.

Now we have the correspondence $n_i \leftrightarrow \lambda_{(i)}$, $i = 1, 2, 3$, while $u \leftrightarrow \lambda_{(0)}$. By (7.53), we also have

$$a_{0i} = -cn_i \cdot \dot{u}\,, \tag{7.94}$$

which means that

$$c\dot{u} = a_{0i}n_i\,, \tag{7.95}$$

since $\dot{u}$ is orthogonal to u.

So, in the notation of Sect. 7.6, which was a completely general construction using any smoothly chosen tetrad along the worldline, and for an arbitrary smooth timelike worldline, the relation

$$c\frac{\mathrm{d}\lambda^\mu_{(\kappa)}}{\mathrm{d}\tau} = \lambda^\mu_{(\nu)}\tilde{A}^{(\nu)}{}_{(\kappa)} \tag{7.96}$$

is replaced by

$$\begin{cases} c\dot{\lambda}_{(i)} = a_{0i}\lambda_{(0)} + c\Omega_{ij}\lambda_{(j)}\,, \\ c\dot{\lambda}_{(0)} = a_{0i}\lambda_{(i)}\,. \end{cases} \tag{7.97}$$

We can now read off the matrix $\tilde{A}$, obtaining

$$\tilde{A}^{(\nu)}{}_{(\kappa)} = \begin{pmatrix} 0 & a_{01} & a_{02} & a_{03} \\ a_{01} & 0 & c\Omega_{21} & c\Omega_{31} \\ a_{02} & c\Omega_{12} & 0 & c\Omega_{32} \\ a_{03} & c\Omega_{13} & c\Omega_{23} & 0 \end{pmatrix}, \tag{7.98}$$

with ν specifying the row and κ the column. Note in passing that, when $\lambda(\tau)$ is obtained by the isometric transport (7.82), namely,

$$\lambda(\tau) = \exp\left(\frac{A\tau}{c}\right)\hat{\lambda}\,, \tag{7.99}$$

we get the same result for $\tilde{A}$ no matter what $\mathrm{ICIF}(\tau) =: K_\tau$ is used to reexpress A, since

$$\begin{aligned} \lambda^{-1}(\tau)A\lambda(\tau) &= \hat{\lambda}^{-1}\exp\left(\frac{-A\tau}{c}\right)A\exp\left(\frac{A\tau}{c}\right)\hat{\lambda} \\ &= \hat{\lambda}^{-1}A\hat{\lambda} = \tilde{A}\,. \end{aligned} \tag{7.100}$$

Returning to the above identification of the matrix $\tilde{A}$ with the matrix on the right-hand side of (7.98), we can immediately deduce what we need to know here in order to prove that we have another rigid frame by this construction, despite the evident fact that the initial tetrad need not be FW transported along the worldline, since we are not assuming $\Omega_{ij} = 0$ for all $i, j \in \{1, 2, 3\}$. The point is that A is a constant matrix if and only if $\tilde{A}$ is a constant matrix, and this is true if and only if a_{0i} and Ω_{ij} are constant for all $i, j \in \{1, 2, 3\}$. This corresponds exactly to superhelical motion as introduced in Sect. 7.6.

At least, we have shown that the theory of GUA in [7.5] always leads to cases of superhelical motion, but it is not yet entirely clear that superhelical motion always corresponds to a case of GUA with isometrically transported triad. After all, if we begin with the relations (7.97), we obtain a relation like (7.96), namely,

$$c\frac{\mathrm{d}\lambda^\mu_{(\kappa)}}{\mathrm{d}\tau} = \lambda^\mu_{(\nu)}\underline{A}^{(\nu)}{}_{(\kappa)}\,, \tag{7.101}$$

where $\underline{A}$ is the constant matrix

$$\underline{A}^{(\nu)}{}_{(\kappa)} := \begin{pmatrix} 0 & a_{01} & a_{02} & a_{03} \\ a_{01} & 0 & c\Omega_{21} & c\Omega_{31} \\ a_{02} & c\Omega_{12} & 0 & c\Omega_{32} \\ a_{03} & c\Omega_{13} & c\Omega_{23} & 0 \end{pmatrix} , \quad (7.102)$$

with ν specifying the row and κ the column, but we have not said anything about how the space triad should be propagated along the worldline. Superhelical motion occurs when the Ω_{ij} are not necessarily zero, but all the a_{0i} and Ω_{ij} are constant, but to show that we have GUA according to the definition, we need to show that we have

$$c\frac{\mathrm{d}u^\mu}{\mathrm{d}\tau} = A^\mu{}_\nu u^\nu , \quad (7.103)$$

for some constant matrix A and for some choice of inertial frame, and we also need to know that the space triad $\{\hat{\lambda}_{(i)}\}_{i=1,2,3}$ has been transported isometrically according to the rule

$$c\frac{\mathrm{d}\lambda^\mu_{(i)}}{\mathrm{d}\tau} = A^\mu{}_\nu \lambda^\nu_{(i)} , \quad i = 1,2,3 . \quad (7.104)$$

This needs to be carefully considered if we are to claim that superhelical motion corresponds precisely to the general GUA construction of Friedman and Scarr.

We can see how to carry out this construction. Starting with (7.101) and (7.102), we have the solution

$$\lambda^\mu_{(\kappa)}(\tau) = \lambda^\mu_{(\nu)}(0)\left[\exp\left(\frac{\underline{A}\tau}{c}\right)\right]^\nu_{(\kappa)} , \quad (7.105)$$

and we define $\hat{\lambda}^\mu_{(\nu)} := \lambda^\mu_{(\nu)}(0)$, which is basically the initial ICIF, whence

$$\lambda^\mu_{(\kappa)}(\tau) = \hat{\lambda}^\mu_{(\nu)}\left[\exp\left(\frac{\underline{A}\tau}{c}\right)\right]^\nu_{(\kappa)} . \quad (7.106)$$

Since we expect $\underline{A}$ to correspond to the matrix $\tilde{A}$ in our previous discussion, we now know how we must define the matrix A by looking at (7.87) and (7.88)

$$A := \hat{\lambda}\underline{A}\hat{\lambda}^{-1} , \quad (7.107)$$

or in component form

$$A^\nu{}_\sigma := \hat{\lambda}^\nu_{(\gamma)}\underline{A}^{(\gamma)}{}_{(\kappa)}(\hat{\lambda}^{-1})^{(\kappa)}{}_\sigma . \quad (7.108)$$

Note that A is constant, i. e., independent of τ, because the matrix $\hat{\lambda}$ is independent of τ. Now what we hope is that

$$c\frac{\mathrm{d}u^\mu}{\mathrm{d}\tau} = A^\mu{}_\nu u^\nu , \quad (7.109)$$

and that $\{\lambda_{(i)}\}_{i=1,2,3}$ is obtained by isometric transport, i. e., by solving

$$c\frac{\mathrm{d}\lambda^\mu_{(i)}}{\mathrm{d}\tau} = A^\mu{}_\nu \lambda^\nu_{(i)} , \quad i = 1,2,3 . \quad (7.110)$$

Since $u = \lambda_{(0)}$, satisfying the last two equations amounts to satisfying

$$c\frac{\mathrm{d}\lambda^\mu_{(\kappa)}}{\mathrm{d}\tau} = A^\mu{}_\nu \lambda^\nu_{(\kappa)} , \quad \kappa = 0,1,2,3 . \quad (7.111)$$

Let us drop indices and work from (7.106) and (7.107) in the form

$$\lambda = \hat{\lambda}\exp\left(\frac{\underline{A}\tau}{c}\right) , \quad A = \hat{\lambda}\underline{A}\hat{\lambda}^{-1} . \quad (7.112)$$

These imply

$$\begin{aligned}\lambda\hat{\lambda}^{-1} &= \hat{\lambda}\exp\left(\frac{\underline{A}\tau}{c}\right)\hat{\lambda}^{-1} \\ &= \exp(\hat{\lambda}\underline{A}\hat{\lambda}^{-1}\tau/c) \\ &= \exp\left(\frac{A\tau}{c}\right) ,\end{aligned}$$

whence

$$\lambda = \exp\left(\frac{A\tau}{c}\right)\hat{\lambda} , \quad (7.113)$$

and this differentiates to give the required result

$$c\frac{\mathrm{d}\lambda}{\mathrm{d}\tau} = A\lambda .$$

Note that the above argument generalizes in a certain sense to nonconstant acceleration tensors.

In the present case, the conclusion from this is that the rigid motion described earlier as superhelical is precisely the motion of fixed space points in the coordinate system constructed by Friedman and Scarr for an observer with GUA in the case where the rotational part of the acceleration matrix is not zero.

Note finally that we do expect the type (2,0) object $\tilde{A}_{(\kappa)(\nu)}$ to be antisymmetric, since

$$\begin{aligned}\lambda_{(\kappa)}\cdot\lambda_{(\nu)} = \eta_{\kappa\nu} &\Rightarrow \dot{\lambda}_{(\kappa)}\cdot\lambda_{(\nu)} + \lambda_{(\kappa)}\cdot\dot{\lambda}_{(\nu)} = 0 \\ &\Rightarrow \left[\lambda_{(\mu)}\tilde{A}^{(\mu)}{}_{(\kappa)}\right]\cdot\lambda_{(\nu)} \\ &\quad + \lambda_{(\kappa)}\cdot\left[\lambda_{(\mu)}\tilde{A}^{(\mu)}{}_{(\nu)}\right] = 0 \\ &\Rightarrow \eta_{\mu\nu}\tilde{A}^{(\mu)}{}_{(\kappa)} + \eta_{\kappa\mu}\tilde{A}^{(\mu)}{}_{(\nu)} = 0 \\ &\Rightarrow \tilde{A}_{(\nu)(\kappa)} + \tilde{A}_{(\kappa)(\nu)} = 0 .\end{aligned} \quad (7.114)$$

The type (2,0) object $\tilde{A}_{(\kappa)(\nu)} := \eta_{\kappa\mu}\tilde{A}^{(\mu)}{}_{(\nu)}$ we obtain from (7.98) is

$$\tilde{A}_{(\kappa)(\nu)} = \begin{pmatrix} 0 & \boldsymbol{a}_0^{\mathrm{T}} \\ -\boldsymbol{a}_0 & c\boldsymbol{\Omega} \end{pmatrix}, \tag{7.115}$$

where κ labels rows and ν labels columns, and

$$\boldsymbol{\Omega} := \begin{pmatrix} \Omega_{11} & \Omega_{12} & \Omega_{13} \\ \Omega_{21} & \Omega_{22} & \Omega_{23} \\ \Omega_{31} & \Omega_{32} & \Omega_{33} \end{pmatrix}, \tag{7.116}$$

and this matrix $\tilde{A}_{(\kappa)(\nu)}$ is indeed antisymmetric.

7.9.6 Summary

Generalized uniform acceleration is a motion satisfying

$$\boxed{c\frac{\mathrm{d}u^{\mu}}{\mathrm{d}\tau} = A^{\mu}{}_{\nu}u^{\nu}}, \tag{7.117}$$

with some specified initial value $u(0) = u_0$, where $A^{\mu}{}_{\nu}$ is a tensor under Lorentz transformations with the property that $A_{\mu\nu} := \eta_{\mu\sigma}A^{\sigma}{}_{\nu}$ is antisymmetric and with the further crucial property of being independent of τ. The initial ICIF $\hat{\lambda} = \{\hat{\lambda}_{(\kappa)}\}$ is transported along the worldline by the isometry specified by

$$\boxed{c\frac{\mathrm{d}\lambda^{\mu}_{(\kappa)}}{\mathrm{d}\tau} = A^{\mu}{}_{\nu}\lambda^{\nu}_{(\kappa)}\,, \quad \lambda_{(\kappa)}(0) = \hat{\lambda}_{(\kappa)}}\,. \tag{7.118}$$

This GUA and the associated semi-Euclidean frame are Poincaré covariant constructions.

When the acceleration matrix A has the translational form

$$A = \begin{pmatrix} 0 & \boldsymbol{g}^{\mathrm{T}} \\ \boldsymbol{g} & 0 \end{pmatrix}, \tag{7.119}$$

where $\boldsymbol{g}$ is constant (independent of τ along the worldline), in some inertial frame relative to which the worldline comes to rest at some event, then the motion is pure TUA according to the standard definition of uniform acceleration and it is straightforward to show that isometric transport (7.118) coincides with FW transport.

So we can construct SE coordinate systems for any observer motion and any smooth propagation of the space triad, but in general, a fluid whose particles sit at fixed space coordinates in such a system will have rigid motion if and only if the space triad is FW transported along the observer worldline. However, for GUA motion with Friedman–Scarr (FS) isometric transport of the space triad, a fluid whose particles sit at fixed space coordinates will also have rigid (superhelical) motion. Indeed such superhelical motion can only be achieved for GUA motion of the main observer and FS transport of the space triad.

7.9.7 Metric for Friedman–Scarr Coordinates

These coordinates are obtained by transporting a tetrad from some initial point on the observer worldline to all other points along it and then carrying out the general construction for an SE coordinate frame. We can thus use the general theory developed earlier, and which leads to the metric form (7.66). We begin with the matrix $\tilde{A}$ given in (7.98), namely,

$$\tilde{A}^{(\nu)}{}_{(\kappa)} = \begin{pmatrix} 0 & a_{01} & a_{02} & a_{03} \\ a_{01} & 0 & c\Omega_{21} & c\Omega_{31} \\ a_{02} & c\Omega_{12} & 0 & c\Omega_{32} \\ a_{03} & c\Omega_{13} & c\Omega_{23} & 0 \end{pmatrix}, \tag{7.120}$$

where ν specifies the row and κ the column. Then relative to the coordinates $\{y^{(\kappa)}\}_{\kappa=0,1,2,3}$, the metric takes the form

$$g_{(\mu)(\nu)} = \begin{pmatrix} \left[1 + y^{(i)}a_{0i}\right]^2 - y^{(i)}y^{(j)}\Omega_{ik}\Omega_{jk} & y^{(i)}\Omega_{i1} & y^{(i)}\Omega_{i2} & y^{(i)}\Omega_{i3} \\ y^{(i)}\Omega_{i1} & -1 & 0 & 0 \\ y^{(i)}\Omega_{i2} & 0 & -1 & 0 \\ y^{(i)}\Omega_{i3} & 0 & 0 & -1 \end{pmatrix} \tag{7.121}$$

Note how this matrix is always independent of the temporal coordinate $y^{(0)}$, and the a_{0i} and Ω_{ij} are just temporal constants for GUA.

This is enough to conclude something that is often (mistakenly) considered important for discussions of the physical interpretation of such coordinate frames, namely that $\partial_{y^{(0)}}$ is a Killing vector field for every such coordinate construction for GUA motion. A Killing vector field X is one such that the Lie derivative $L_X g$ of the metric along the flow curves of X is zero.

To prove this claim, we may use the general coordinate formula for the Lie derivative as given in [7.6]. For

any contravariant vector field X, we have

$$(L_X g)_{(\eta)(\phi)} = \frac{\partial g_{(\eta)(\phi)}}{\partial y^{(\iota)}} X^{(\iota)} + g_{(\iota)(\phi)} \frac{\partial X^{(\iota)}}{\partial y^{(\eta)}} + g_{(\eta)(\iota)} \frac{\partial X^{(\iota)}}{\partial y^{(\phi)}} . \tag{7.122}$$

We then take $X = \partial_{y^{(0)}}$ which has components $X^{(0)} = 1$, $X^{(i)} = 0$, $i = 1,2,3$, in these coordinates. Hence,

$$(L_X g)_{(\eta)(\phi)} = \frac{\partial g_{(\eta)(\phi)}}{\partial y^{(0)}} X^{(0)} = 0 ,$$

as claimed. We can thus say that all observers sitting at fixed space coordinate positions in these frames are Killing observers.

Any spacetime with a metric of the form (7.121) has a globally defined timelike Killing vector field and is said to be stationary. If in addition only the diagonal elements are nonzero, as happens when all the Ω_{ij} are zero and we have TUA, the spacetime is said to be static. Of course, this is the flat Minkowski spacetime so we already know that it is static. What we discover here is the plethora of Killing vector fields that can be used to get the Minkowski metric into the stationary or static forms.

7.9.8 More about Observers at Fixed Space Coordinates

A more general question is whether these Killing observers sitting at fixed space coordinates in the $\{y^{(\kappa)}\}_{\kappa=0,1,2,3}$ system actually have GUA motion. In order to tackle this, we need to know the proper time of these observers.

Here we can also use the general theory of semi-Euclidean coordinate systems in Sect. 7.6. Recall that this analysis considers a space triad $\{n_i\}_{i=1,2,3}$ that is smoothly transported along the observer worldline, without assuming anything other than smoothness about the transport. Furthermore, we have made the link with the quantities a_{0i} and Ω_{ij} in the relations (7.93), namely,

$$c\dot{n}_i^{\ \mu} = a_{0i} u^\mu + c\Omega_{ij} n_j^{\ \mu} . \tag{7.123}$$

So as we saw previously, we have the correspondence $n_i \leftrightarrow \lambda_{(i)}$, $i = 1, 2, 3$, while $u \leftrightarrow \lambda_{(0)}$ and

$$a_{0i} = -cn_i \cdot \dot{u} , \tag{7.124}$$

which means that

$$c\dot{u} = a_{0i} n_i , \tag{7.125}$$

since $\dot{u}$ is orthogonal to u. Then, in this notation, which was a completely general construction using any smoothly chosen tetrad along the worldline, and for an arbitrary smooth timelike worldline, the relation

$$c \frac{\mathrm{d}\lambda^\mu_{(\kappa)}}{\mathrm{d}\tau} = \lambda^\mu_{(\nu)} \tilde{A}^{(\nu)}{}_{(\kappa)}$$

is replaced by

$$\begin{cases} c\dot{\lambda}_{(i)} = a_{0i}\lambda_{(0)} + c\Omega_{ij}\lambda_{(j)} , \\ c\dot{\lambda}_{(0)} = a_{0i}\lambda_{(i)} , \end{cases} \tag{7.126}$$

and we read off the matrix $\tilde{A}$ as

$$\tilde{A}^{(\nu)}{}_{(\kappa)} = \begin{pmatrix} 0 & a_{01} & a_{02} & a_{03} \\ a_{01} & 0 & c\Omega_{21} & c\Omega_{31} \\ a_{02} & c\Omega_{12} & 0 & c\Omega_{32} \\ a_{03} & c\Omega_{13} & c\Omega_{23} & 0 \end{pmatrix} , \tag{7.127}$$

with ν specifying the row and κ the column. The specific feature of GUA motion is that the a_{0i} and Ω_{ij} are actually independent of the proper time along the observer worldline. It is also important to note that the point of contact between this analysis and Friedman and Scarr's is through $\tilde{A}$, the expression for the acceleration matrix relative to any ICIF for the main observer, rather than through A, the expression for the acceleration matrix relative to some arbitrary laboratory inertial frame.

Now it is established in Sect. 7.6 that the 4-velocity of an observer sitting at fixed $\xi^i \leftrightarrow y^{(i)}$ in the accelerating frame is (see (7.54))

$$u^\mu(\xi,\tau) = \left[\left(1 + \xi^i a_{0i}\right) u_0{}^\mu + \xi^i \Omega_{ij} n_j{}^\mu\right] \dot{\sigma} , \tag{7.128}$$

where τ is the proper time for the observer at ξ and $\sigma(\xi,\tau)$ is the corresponding proper time of the main observer, corresponding in the sense that, at that proper time, the main observer considers the observer at ξ to be simultaneous. The dot on σ denotes the derivative with respect to τ, keeping ξ fixed, so it is the time dilation effect between the two observers. In fact, it was shown in (7.55) that

$$\dot{\sigma} = \left[\left(1 + \xi^i a_{0i}\right)^2 - \xi^i \xi^j \Omega_{ik} \Omega_{jk}\right]^{-1/2} . \tag{7.129}$$

Note that $\dot{\sigma}$ is constant for GUA motion, because then a_{0i} and Ω_{ij} are constant and we have fixed the ξ^i. So the full formula for the 4-velocity of the observer sitting at fixed ξ is

$$u^\mu(\xi,\tau) = \frac{\left(1+\xi^i a_{0i}\right) u_0{}^\mu + \xi^i \Omega_{ij} n_j{}^\mu}{\left[(1+\xi^i a_{0i})^2 - \xi^i \xi^j \Omega_{ik}\Omega_{jk}\right]^{1/2}} . \tag{7.130}$$

We must now obtain the 4-acceleration $a^\mu(\xi,\tau)$ by differentiating $u^\mu(\xi,\tau)$ with respect to τ for fixed ξ. The aim will be to see whether the 4-acceleration can be obtained by multiplying the 4-velocity by some constant matrix. We have

$$\begin{aligned} a^\mu(\xi,\tau) &= \left.\frac{\partial u^\mu(\xi,\tau)}{\partial \tau}\right|_\xi \\ &= \frac{\left(1+\xi^i a_{0i}\right) \dot{u}_0{}^\mu + \xi^i \Omega_{ij} \dot{n}_j{}^\mu}{\left[(1+\xi^i a_{0i})^2 - \xi^i \xi^j \Omega_{ik}\Omega_{jk}\right]^{1/2}}\dot{\sigma} \\ &= \frac{\left(1+\xi^i a_{0i}\right) A^\mu{}_\nu u_0{}^\nu + \xi^i \Omega_{ij} A^\mu{}_\nu n_j{}^\nu}{(1+\xi^i a_{0i})^2 - \xi^i \xi^j \Omega_{ik}\Omega_{jk}} \\ &= \frac{A^\mu{}_\nu}{\left[(1+\xi^i a_{0i})^2 - \xi^i \xi^j \Omega_{ik}\Omega_{jk}\right]^{1/2}} \\ &\quad \times u^\nu(\xi,\tau) , \end{aligned} \tag{7.131}$$

using the fact that $\dot{u}_0{}^\mu = A^\mu{}_\nu u_0{}^\nu$ and $\dot{n}_j{}^\mu = A^\mu{}_\nu n_j{}^\nu$, where $A^\mu{}_\nu$ is the version of the constant acceleration matrix expressed relative to the laboratory inertial frame. We conclude that an observer sitting at fixed ξ^i in the FS accelerating frame would indeed have GUA, with the acceleration matrix

$$\boxed{A^\mu{}_\nu(\xi) = \frac{A^\mu{}_\nu}{\left[(1+\xi^i a_{0i})^2 - \xi^i \xi^j \Omega_{ik}\Omega_{jk}\right]^{1/2}}} , \tag{7.132}$$

as expressed relative to the laboratory inertial frame. When the latter is the initial instantaneously comoving inertial frame ICIF(0) of the main observer, or in fact any ICIF of the main observer and hence also of the observer at ξ, we have

$$\boxed{\begin{aligned}\tilde{A}^\mu{}_\nu(\xi) &= \frac{1}{\left[(1+\xi^i a_{0i})^2 - \xi^i \xi^j \Omega_{ik}\Omega_{jk}\right]^{1/2}} \\ &\quad\times \begin{pmatrix} 0 & a_{01} & a_{02} & a_{03} \\ a_{01} & 0 & \Omega_{21} & \Omega_{31} \\ a_{02} & \Omega_{12} & 0 & \Omega_{32} \\ a_{03} & \Omega_{13} & \Omega_{23} & 0 \end{pmatrix}\end{aligned}} , \tag{7.133}$$

although it is not necessary to see this form in order to prove the above claim.

It can be shown that observers sitting at fixed space coordinates in the FS frame of a main observer with GUA motion share hyperplanes of simultaneity with the latter, in the precise sense described just after (7.63), if and only if the motion of the main observer is actually TUA. But in fact HOS sharing also occurs for a main observer with arbitrary motion provided she uses an FW transported tetrad to establish coordinates, regardless of whether her purely translational acceleration as viewed in this frame is uniform or not.

7.10 A Brief Conclusion

There is an important difference between inertial frames and what we have been referring to as accelerating frames: when an observer has inertial motion, either in special relativity or in general relativity, we know what is the most *natural* frame for such a person to use, namely, an inertial or locally inertial frame, since it is in these frames that our theories of nongravitational physics take on their simplest forms. This in turn is ultimately related to the fact that the latter theories, which govern whatever is being measured and whatever is being used to measure them, have a velocity symmetry, which we usually refer to as Lorentz or local Lorentz symmetry, respectively.

But when the observer is moving with some acceleration, although we may still find adapted frames in the sense outlined back at the beginning of Sect. 7.2, we cannot simple-mindedly pretend that our theories of physics expressed relative to such frames look just as

they would in inertial frames. We must remember that our accelerating frames merely provide us with a coordinate description that happens to be convenient in some ways. This in turn is ultimately related to the fact that our theories of nongravitational physics have no acceleration symmetries, at least as far as we know. For in-depth discussion of this problem, see [7.3].

It should also be remembered that what we have referred to as rigid motion is very much a theoretical notion. Whether it could ever be achieved by any continuous material is quite another question [7.7]. In fact, the whole of this chapter is concerned primarily with mathematical aspects of our modern spacetime theories. Their physical interpretation is another matter.

References

7.1 W. Rindler: *Introduction to Special Relativity* (Oxford Univ. Press, New York 1982)

7.2 M. Friedman: *Foundations of Space-Time Theories* (Princeton Univ. Press, Princeton 1983)

7.3 S.N. Lyle: *Uniformly Accelerating Charged Particles. A Threat to the Equivalence Principle*, Fundamental Theories of Physics, Vol. 158 (Springer, Berlin Heidelberg 2008), see in particular Chap. 2

7.4 B. DeWitt: *Bryce DeWitt's Lectures on Gravitation* (Springer-Verlag, Berlin Heidelberg 2011), The notation and formulation here are very largely inspired by these lecture notes

7.5 Y. Friedman, T. Scarr: Covariant uniform acceleration (2011) arXiv:1105.0492v2 [phys.gen-ph]

7.6 S.W. Hawking, G.F.R. Ellis: *The Large Scale Structure of Space-Time* (Cambridge Univ. Press, Cambridge 1973)

7.7 S.N. Lyle: *Self-Force and Inertia. Old Light on New Ideas*, Lecture Notes in Physics, Vol. 796 (Springer, Berlin Heidelberg 2010), Chap. 12

8. Physics as Spacetime Geometry

Vesselin Petkov

As there have been no major advancements in fundamental physics in the past decades it seems reasonable to reexamine the major explicit and especially implicit assumptions in fundamental physics to ensure that all logically possible research directions are identified. The purpose of this chapter is to outline such a direction. Minkowski's program of regarding four-dimensional physics as spacetime geometry is rigorously and consistently employed to the already geometrized general relativity with the most stunning implication that gravitational phenomena are fully explained in the theory without the need to assume that they are caused by gravitational interaction. Then the real open question in gravitational physics seems to be how matter curves spacetime, not how to quantize the apparent gravitational interaction. In view of the difficulties encountered by quantum gravity, even the radical option that gravity is not a physical interaction deserves careful scrutiny due to its potential impact on fundamental physics as a whole. The chapter discusses the possible implications of this option for the physics of gravitational waves and for quantum gravity and ends with an example where regarding physics as spacetime geometry provides a straightforward explanation of a rather subtle issue in relativity – propagation of light in noninertial reference frames.

8.1 Foundational Knowledge and Reality of Spacetime

Minkowski's program of regarding four-dimensional physics as spacetime geometry is often viewed as just a more convenient description of physical phenomena. However, I think Minkowski's program is crucially important for fundamental physics; hence, the program and its implications should be rigorously examined for the following reason. The identification of four-dimensional physics with the geometry of spacetime presupposes that spacetime represents a real four-dimensional world as Minkowski insisted since physics cannot be geometry of something abstract (here we again face the challenging question of whether a mathematical formalism is only a convenient description of physical phenomena or reveals true features of the physical world). However, the status of spacetime has been unresolved and this might turn out to be ultimately re-

sponsible for the failure to create a quantum theory of gravity so far, and possibly even for the fact that in the last several decades there has been no major breakthrough as revolutionary as the theory of relativity and quantum mechanics despite the unprecedented advancements in applied physics and technology and despite the efforts of many brilliant physicists.

It is not inconceivable to assume that the present state of fundamental physics may be caused by some metatheoretical problems, not by the lack of sufficient experimental evidence and talented physicists. I think the major metatheoretical reason for most difficulties in contemporary fundamental physics, and particularly for not dealing with the status of spacetime, is underestimating the necessity to identify explicitly which elements of our theories adequately represent elements of the physical world. Such reliable knowledge about the world is a necessary condition for the smooth advancement of fundamental physics since it forms the foundation on which new theories are built. To ensure that such foundational knowledge will never be revised as our understanding of the world deepens, knowledge should be rigorously and *unambiguously* extracted from the *experimental* evidence. As experiments do not contradict one another no future discoveries can challenge the accumulated foundational knowledge. In 1909 *Max Planck* expressed the idea of foundational knowledge (whose elements he properly called invariants) perhaps in the best possible way [8.1]:

> *The principle of relativity holds not only for processes in physics but also for the physicist himself, in that a fixed system of physics exists in reality only for a given physicist and for a given time. But, as in the theory of relativity, there exist invariants in the system of physics: ideas and laws which retain their meaning for all investigators and for all times, and to discover these invariants is always the real endeavor of physical research. We shall work further in this direction in order to leave behind for our successors where possible – lasting results. For if, while engaged in body and mind in patient and often modest individual endeavor, one thought strengthens and supports us, it is this, that we in physics work, not for the day only and for immediate results, but, so to speak, for eternity.*

In close connection with the necessity for explicit foundational knowledge, it is worth stressing that a view, which some physicists are sometimes tempted to hold – that physical phenomena can be described *equally* by different theories (*it is just a matter of description*) – hampers our understanding of the world and negatively affects the advancement of fundamental physics since such a view effectively rules out the need for foundational knowledge. I hope all will agree that part of the art of doing physics is to determine whether different theories are indeed simply different descriptions of the same physical phenomena (as is the case with the three representations of classical mechanics – Newtonian, Lagrangian, and Hamiltonian), or *only one* of the theories competing to describe and explain given physical phenomena is the correct one (as is the case with general relativity, which identifies gravity with the non-Euclidean geometry of spacetime, and other theories, which regard gravity as a force).

Due to the unsettled status of spacetime, there are physicists who hold the experimentally unsupported view that the concept of spacetime is only a successful *description* of the world (an *abstract bookkeeping structure* [8.2]) and for this reason it is nothing more than *an abstract four-dimensional mathematical continuum* [8.2]. Therefore, on this view, the concept of spacetime does not imply *that we inhabit a world that is such a four- (or, for some of us, ten-) dimensional continuum* [8.2]. In addition to not being backed by experiment, the problem with this view is that it is an unproductive one since it makes it impossible even to identify what the implications of a real spacetime are. As those implications might turn out to be necessary for the advancement of fundamental physics, Sect. 8.2 deals with the essence of the spacetime concept (the reality of spacetime, i. e., that the world is four-dimensional) and argues that the relativistic experimental evidence provides strong support for it, which allows us to regard the reality of spacetime as an important piece of foundational knowledge. Section 8.2 also examines how Minkowski's program of geometrization of physics sheds additional light on Einstein's geometrization of gravity and suggests that gravitational phenomena are not caused by gravitational interaction since those phenomena are fully explained in general relativity without the need of gravitational interaction. The implications of this possibility for the search for gravitational waves and for quantum gravity are discussed in the last part of the section. Section 8.3 demonstrates how taking the reality of spacetime explicitly into account makes it self-evident why the propagation of light in noninertial reference frames in flat and curved spacetimes is anisotropic. (Strictly speaking, the expression *propagation of light in flat and curved spacetime* is incorrect. Nothing propagates

or moves in spacetime since the whole history of every particle is entirely given (at once) as the particle's worldline in spacetime. Such expressions only mean to state how the null worldlines of light rays in flat and curved spacetime are expressed in the ordinary three-dimensional language as propagation of those rays.)

8.2 Four-Dimensional Physics as Spacetime Geometry

In the beginning of this section I will summarize what I think is the unequivocal experimental evidence which indicates that the concept of spacetime does represent a real four-dimensional world. If the arguments convincingly show (as I believe they do) that *the relativistic experimental evidence would be impossible if the world were three dimensional*, then the reality of spacetime (i. e., the assertion that the world is four-dimensional) is indeed a major piece of foundational knowledge.

It was Minkowski who initially extracted this foundational knowledge from the experimental evidence that supported the relativity principle. On September 21, 1908 he began his famous lecture *Space and Time* by announcing the revolutionary view of space and time, which he deduced from experimental physics by successfully decoding the profound message hidden in the failed experiments to discover absolute motion [8.3, p.111]:

> *The views of space and time which I want to present to you arose from the domain of experimental physics, and therein lies their strength. Their tendency is radical. From now onwards space by itself and time by itself will recede completely to become mere shadows and only a type of union of the two will still stand independently on its own.*

Minkowski repeatedly stressed the *experimental* fact that absolute motion and absolute rest cannot be discovered:

> *All efforts directed towards this goal, especially a famous interference experiment of Michelson had, however, a negative result.* [8.3, p. 116].
>
> *In light of Michelson's experiment, it has been shown that, as Einstein so succinctly expresses this, the concept of an absolute state of rest entails no properties that correspond to phenomena.* [8.4].

Minkowski had apparently felt that the experimental evidence supporting Galileo's principle of relativity (absolute motion with constant velocity cannot be discovered through mechanical experiments) and the failed experiments (involving light beams) to detect the Earth's motion contained some hidden information about the physical world that needed to be decoded. That is why he had not been satisfied with the principle of relativity which merely *postulated* that absolute motion and absolute rest did not exist. To decode the hidden information, *Minkowski* first examined (as a mathematician) the fact that *The equations of Newtonian mechanics show a twofold invariance* [8.3, p. 111]. As each of the two invariances represents a certain group of transformations for the differential equations of mechanics Minkowski noticed that the second group (representing invariance with respect to uniform translations, i. e., Galileo's principle of relativity) leads to the conclusion that the *time axis can then be given a completely arbitrary direction in the upper half of the world* $t > 0$. This strange implication made *Minkowski* ask the question that led to the new view of space and time [8.3, p. 111]:

> *What has now the requirement of orthogonality in space to do with this complete freedom of choice of the direction of the time axis upwards?*

In answering this question Minkowski showed *why* the time t of a stationary observer and the time t', which *Lorentz* introduced (as *an auxiliary mathematical quantity* [8.5]) calling it the *local time* of a moving observer (whose x'-axis is along the x-axis of the stationary observer), should be treated equally (which Einstein simply *postulated* in his 1905 paper) [8.3, p. 114]:

> *One can call t' time, but then must necessarily, in connection with this, define space by the manifold of three parameters x', y, z in which the laws of physics would then have exactly the same expressions by means of x', y, z, t' as by means of x, y, z, t. Hereafter we would then have in the world no more* the *space, but an infinite number of spaces analogously as there is an infinite number of planes in three-dimensional space. Three-dimensional geometry becomes a chapter in four-dimensional physics. You see why I said at the beginning that space and time will recede completely to become mere shadows and only a world in itself will exist.*

The profound implication of *the requirement of orthogonality in space* is evident in the beginning of this quote – as t and t' are two different times it necessar-

ily follows that two different spaces must be associated with these times since each space is orthogonal to each time axis. *Minkowski* easily saw the obvious for a mathematician fact that different time axes imply different spaces and remarked that *the concept of space was shaken neither by Einstein nor by Lorentz* [8.3]. Then, as the quote demonstrates, *Minkowski* had immediately realized that many spaces and times imply that the world is four-dimensional with *all* moments of time forming the fourth dimension (Poincaré showed before Minkowski that the Lorentz transformations can be regarded as rotations in a four-dimensional space with time as the fourth axis but, unlike Minkowski, he did not believe that such a four-dimensional mathematical space represented anything in the world; see the Introduction of [8.6], particularly pages 19–23, and the reference therein).

Minkowski excitedly announced the new views of space and time since he clearly recognized that their strength comes from the fact that they *arose from the domain of experimental physics* – the arguments that many times imply many spaces as well, which in turn implies that the world is four-dimensional, are deduced unambiguously from the *experiments* that confirmed the principle of relativity (i. e., the impossibility to discover absolute uniform motion and absolute rest). Indeed, all physical phenomena look in the same way to two observers A and B in uniform relative motion (so they cannot tell who is moving as the experimental evidence proved) *because* A and B have different times (as Lorentz formally proposed, Einstein postulated and Minkowski explained) and different spaces – each observer performs experiments in his own space and time and for this reason the physical phenomena look in the same way to A and B (e.g., the speed of light is the same for them since each observer measures it in his own space by using his own time). This explanation of the profound meaning of the principle of relativity, extracted from experimental physics, makes the nonexistence of absolute motion and absolute rest quite evident – absolute motion and absolute rest do not exist since they are defined with respect to an absolute (single) space, but such a single space does not exist in the world; all observers in relative motion have their own spaces and times.

The most direct way to evaluate Minkowski's confidence in the strength of the new views of space and time and his insistence that they were deduced from experimental physics is to assume, for the sake of the argument, that spacetime is nothing more than *an abstract four-dimensional mathematical continuum* [8.2] and that the physical world is three dimensional. Then there would exist a *single* space (since a three-dimensional world presupposes the existence of one space), which as such would be absolute (the same for all observers). As a space constitutes a class of simultaneous events (the space points at a given moment), a single (absolute) space implies absolute simultaneity and therefore absolute time as well. Hence a three-dimensional world allows *only* absolute space and absolute time in contradiction with the experimental evidence that uniform motion with respect to the absolute space cannot be discovered as encapsulated in the principle of relativity. *Minkowski*'s realization that the world must be four-dimensional in order that absolute motion and rest do not exist naturally explains his dissatisfaction with the principle of relativity, which postulates, but does not explain the nonexistence of absolute motion and rest [8.3, p. 117]:

> *I think the word* relativity postulate *used for the requirement of invariance under the group* G_c *is very feeble. Since the meaning of the postulate is that through the phenomena only the four-dimensional world in space and time is given, but the projection in space and in time can still be made with a certain freedom, I want to give this affirmation rather the name* the postulate of the absolute world.

In addition to Minkowski's arguments, I would like to stress what I consider to be a fact that special relativity and particularly the *experiments*, which confirmed the kinematical relativistic effects, are *impossible* in a three-dimensional world. I think each of the arguments listed below taken even alone is sufficient to demonstrate that:

- Relativity of simultaneity is impossible in a three-dimensional world – as a three-dimensional world (like a three-dimensional space) is a class of simultaneous events (everything that exists simultaneously at the present moment), if the physical world were three dimensional, there would exist a single class of simultaneous events; therefore simultaneity would be absolute since all observers in relative motion would share the same three-dimensional world and therefore the same class of simultaneous events.
- Since length contraction and time dilation are specific manifestations of relativity of simultaneity they are also impossible in a three-dimensional world. What is crucial is that the *experiments* which confirmed these relativistic effects would be impossible if the physical world were three dimensional

[8.7, Chap. 5]. Along with time dilation, the muon experiment (see, for example, [8.8, p. 103]) effectively tested length contraction experimentally as well [8.8, p. 104] –

In the muon's reference frame, we reconcile the theoretical and experimental results by use of the length contraction effect, and the experiment serves as a verification of this effect.

- The twin paradox effect and the experiments that confirmed it are also impossible in a three-dimensional world [8.7, Chap. 5].

A valuable concrete example of why special relativity is impossible in a three-dimensional world is Minkowski's explanation of the deep physical meaning of length contraction as depicted in figure 1 of his paper *Space and Time* whose right-hand part is reproduced in Fig. 8.1.

The essence of his explanation (which is the accepted correct explanation) is that the relativistic length contraction of a body is a manifestation of the reality of the body's worldline or rather worldtube (for a spatially extended body). Minkowski considered two bodies in uniform relative motion represented by their worldtubes as shown in Fig. 8.1. To see clearly why *the worldtube of a body must be real in order that length contraction be possible*, consider the body represented by the vertical worldtube. The three-dimensional cross-section PP, resulting from the intersection of the body's worldtube and the space (represented by the horizontal line in Fig. 8.1) of an observer at rest with respect to the body, is the body's proper length. The three-dimensional cross-section $P'P'$, resulting from the intersection of the body's worldtube and the space (represented by the inclined dashed line) of an observer at rest with respect to the second body (represented by

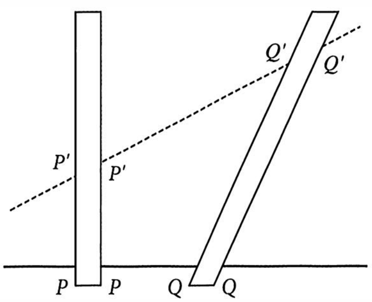


Fig. 8.1 The right-hand part of figure 1 in Minkowski's paper *Space and Time*

the inclined worldtube), is the relativistically contracted length of the body measured by that observer (the cross-section $P'P'$ only appears longer than PP because a fact of the pseudo-Euclidean geometry of spacetime is represented on the Euclidean surface of the page). Note that while measuring the *same* body, the two observers measure *two* three-dimensional bodies represented by the cross-sections PP and $P'P'$ in Fig. 8.1 (this relativistic situation will not be truly paradoxical only if what is meant by *the same body* is the body's worldtube).

In order to judge the argument that length contraction is impossible in a three-dimensional world, assume that the worldtube of the body did not exist as a four-dimensional object and were nothing more than an abstract geometrical construction. Then, what would exist would be a single three-dimensional body, represented by the proper cross-section PP, and both observers would measure the *same* three-dimensional body of the *same* length. Therefore, not only would length contraction be *impossible*, but relativity of simultaneity would be also impossible since a spatially extended three-dimensional object is defined in terms of *simultaneity* – all parts of a body taken *simultaneously* at a given moment – and as both observers in relative motion would measure the same three-dimensional body (represented by the cross-section PP) they would share the *same* class of simultaneous events in contradiction with relativity.

After Minkowski had successfully decoded the profound message hidden in the failed experiments to detect absolute uniform motion and absolute rest – that the world is four-dimensional – he had certainly realized that four-dimensional physics was in fact spacetime geometry since all particles which *appear* to move in space are in reality a forever given web of the particles' worldlines in spacetime. Then *Minkowski* outlined the program of geometrization of physics [8.3, p. 112]:

The whole world presents itself as resolved into such worldlines, and I want to say in advance, that in my understanding the laws of physics can find their most complete expression as interrelations between these worldlines.

And before his tragic and untimely departure from this world on January 12, 1909 he started to implement this program as will be briefly discussed below. But let me first address a view which some physicists are sometimes tempted to hold – that we should not take the implications of special and general relativity too seriously because these theories cannot accommodate the probabilistic behavior of quantum objects.

In fact, it is that view which should not be taken seriously for two reasons. First, as it is the *experiments* confirming the kinematical relativistic effects that would be impossible in a three-dimensional world, the reality of spacetime (the four-dimensionality of the world) must be treated with utmost seriousness. Since *experiments do not contradict one another* no future experiments can force us to abandon the view that the world is four-dimensional and that macroscopic bodies are worldtubes in this world. Second, the fact that elementary particles are not worldlines in spacetime only indicates *what they are not* and in no way tells us something against the reality of spacetime. Elementary particles, or perhaps more appropriately quantum objects, might be more complex structures in spacetime (for a conceivable example see [8.7, Chap. 10] and the references therein). As an illustration that spacetime can accommodate probability perfectly well, imagine that the probabilistic behavior of the quantum object is merely a manifestation of a *probabilistic distribution of the quantum object itself in the forever given spacetime* – an electron, for instance, can be thought of (for the sake of the argument that spacetime structures can be probabilistic as well) as an ensemble of the points of its *disintegrated* worldline which are scattered in the spacetime region where the electron wavefunction is different from zero. Had Minkowski lived longer he might have described such a probabilistic spacetime structure by the mystical expression *predetermined probabilistic phenomena.*

I think the very fact that the status of spacetime has not been firmly settled for over a 100 year deserves special attention since it may provide some valuable lessons for the future of fundamental physics. Indeed, it is *logically* inexplicable why Minkowski's effective arguments for the reality of spacetime have been merely ignored (they have not been disproved); as we saw above his arguments taken alone demonstrate that the world must be four-dimensional in order that special relativity and the experimental evidence which tested its kinematical effects be possible. It appears the reason for ignoring the arguments for the reality of spacetime are not scientific; the reason does not seem to be even rational since those arguments are merely regarded as nonexistent. Quite possibly, such an attitude towards the nature of spacetime is caused by the temptation to regard the claim that the physical world is four-dimensional as an outrageously and self-evidently false, because of the colossally counter-intuitive nature of such a world and because of its huge implications for virtually all aspects of our lives. Perhaps such a reaction to arguments for disturbingly counter-intuitive new discoveries was best shown by Cantor in a letter to Dedekind in 1877 where he commented on the way he viewed one of his own major results (the one-to-one correspondence of the points on a segment of a line with i) the points on an indefinitely long line, ii) the points on a plane, and iii) the points on any multidimensional mathematical space) – *I see it, but I don't believe it* [8.9]. However, the nature of the world as revealed by the experimental evidence – no matter how counter-intuitive it may be – should be faced and should not be squeezed into our preset and deceivingly comfortable views about what exists.

Due to the unsettled status of spacetime so far, Minkowski's program of adequately treating four-dimensional physics as spacetime geometry has not been fully implemented. As a result, new discoveries leading to a deeper understanding of the world might have been delayed. A small example is the propagation of light in noninertial reference frames – this issue could have been addressed and clarified immediately after Minkowski's four-dimensional formulation of special relativity. In the remaining part of this section I will discuss first Minkowski's initial steps of the implementation of his program of geometrization of physics and then will outline other unexplored implications of his program some of which may have significantly affected front line research programs in fundamental physics such as the search for gravitational waves and quantum gravity.

8.2.1 Generalization of Inertial Motion in Special and General Relativity

Minkowski generalized Newton's first law (of inertia) for the case of flat spacetime by noticing that a free particle, which is at (relative) rest or moves by inertia, is a straight timelike worldline. Then he pointed out that an accelerating particle is represented by a curved worldline. Here is how *Minkowski* described the three states of motion of a particle (corresponding to the worldlines a, b, and c in Fig. 8.2) [8.3, p. 115]:

> *A straight worldline parallel to the t-axis corresponds to a stationary substantial point, a straight line inclined to the t-axis corresponds to a uniformly moving substantial point, a somewhat curved worldline corresponds to a nonuniformly moving substantial point.*

As a straight timelike worldline represents inertial motion it immediately becomes clear why experiments

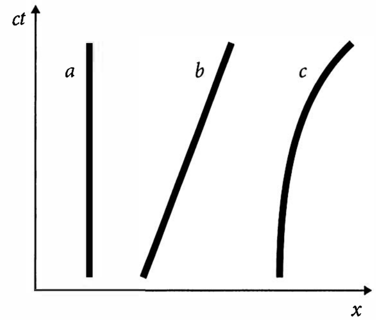


Fig. 8.2 Worldlines a and b represent two particles – one at rest (a) and the other in uniform motion (b), whereas worldline (c) represents an accelerating particle

have always failed to distinguish between a state of rest and a state of uniform motion – in both cases a particle is a *straight* worldline as seen in Fig. 8.2 (worldlines a and b) and there is clearly no distinction between two straight lines. In the figure the time axis of the reference frame is along worldline a and the particle represented by this worldline appears to be at rest in the reference frame. If the time axis of another reference frame is chosen along worldline b, the particle represented by that worldline will appear to be at rest in the new reference frame, whereas the first particle (represented by worldline a) will appear to be uniformly moving with respect to the second particle (since worldline a is inclined to the new time axis, i. e., inclined to worldline b). *Minkowski* seems to have been impressed by this elegant explanation of the experimental fact that rest and uniform motion cannot be distinguished (which is a more detailed *explanation* of the relativity principle) that he decided [8.10, p. 115]

> *to introduce this fundamental axiom:* With appropriate setting of space and time the substance existing at any worldpoint can always be regarded as being at rest.

Perhaps the most successful continuation of Minkowski's program of geometrization of physics is the generalization of inertial motion in general relativity. This is encapsulated in the geodesic hypothesis in general relativity, which states that the worldline of a free particle is a timelike *geodesic* in spacetime. This hypothesis is regarded as *a natural generalization of Newton's first law* [8.11, p. 110], that is, *a mere extension of Galileo's law of inertia to curved spacetime* [8.12]. This means that *in general relativity a particle, whose worldline is geodesic, is a free particle which moves by inertia.*

Unfortunately, the important implications of this rare implementation of Minkowski's program have not been fully explored, which might have delayed the research in gravitational physics, particularly the initiation and advancement of a research program to reveal the mechanism of how matter curves spacetime. The immediate consequence of the geodesic hypothesis and its experimental confirmation by the fact that falling bodies do not resist their fall (a falling accelerometer, for example, reads zero resistance, i. e., zero absolute acceleration, since it measures acceleration through resistance) implies that the explanation of gravitational phenomena does not need the assumption of the existence of gravitational interaction. The reason is that as a falling body moves by inertia (since it does not resist its fall) no gravitational force is causing its fall, i. e., it is not subject to any interactions since inertial, i. e., nonresistant, motion by its very nature is *interaction-free* motion. The analysis of this consequence of the geodesic hypothesis naturally leads to the question of how matter curves spacetime in order to determine whether the Earth interacts gravitationally with a falling body through the curvature of spacetime. If there is such an interaction between the Earth and the body, there should exist *extra* stress energy of the Earth not only to curve spacetime but to change the shape of the geodesic worldtube of the falling body (that change of shape makes it more *curved*, but not deformed, as will be discussed below, which means that the worldtube of the falling body is geodesic and the body does not resist its fall). As we will see below this does not appear to be the case since the Einstein–Hilbert equation implies that no extra stress energy is necessary to change the shape of the geodesic worldtube of a falling body – the same stress energy of the Earth, for example, produces the same spacetime curvature no matter whether or not there are other bodies in the Earth's vicinity.

The importance of the experimental fact that falling bodies offer no resistance to their fall is that *it rules out* any alternative theories of gravity and any attempts to quantize gravity (by proposing alternative representations of general relativity aimed at making it amenable to quantization) that regard gravity as a *physical* field which gives rise to a gravitational *force* since they would contradict the experimental evidence. It should be particularly stressed that a gravitational force would be required to move particles downwards *only if* the particles *resisted* their fall, because *only then* a gravitational force would be needed to *overcome* that resistance.

8.2.2 In What Sense Is Acceleration Absolute in Both Special and General Relativity?

Minkowski's own implementation of his program to represent four-dimensional physics as spacetime geometry produced another important result – an unforeseen resolution of the debate over the status of acceleration, which was prompted by Newton's insistence that both acceleration and space are absolute (since acceleration is experimentally detectable, and therefore absolute, which implies that space is also absolute due to the apparently self-evident assumption that any acceleration is with respect to space).

Encouraged by the resolution of the centuries-old puzzle (that it is impossible to distinguish experimentally between rest and uniform motion) in terms of the geometry of spacetime – two particles, one at rest and the other in uniform motion, are both *straight* worldlines in spacetime – *Minkowski* almost certainly had immediately seen that another experimental fact – acceleration is experimentally detectable – also had an elegant explanation in terms of spacetime geometry: an accelerating particle is a *curved* worldline in spacetime. He expressed this observation by stressing that *Especially the concept of* acceleration *acquires a sharply prominent character* [8.3, p. 117].

Minkowski left this world less than 4 months after he gave his last and famous lecture *Space and Time* where he talked about that *sharply prominent character* of acceleration. He was not given the chance to develop further his ideas. But Minkowski succeeded in revealing the deep physical meaning of the distinction between inertial and accelerated motion: the absolute physical facts that inertial motion cannot be detected experimentally, whereas accelerated motion is experimentally detectable, correspond to two geometrical facts in spacetime – a particle moving by inertia is a *straight* worldline, whereas an accelerating particle is a *curved* worldline. Such an explanation of physical facts by facts of the geometry of spacetime is not only natural, but is the only explanation in a real four-dimensional world – in such a world, i. e., in spacetime, there are no three-dimensional particles which move inertially or with an acceleration; there is only a forever given network of straight and curved worldlines there.

The *absoluteness* (frame-independence) of acceleration and inertial motion is reflected in the *curvature* of the worldline of an accelerating particle and the *straightness* of the worldline of a particle moving by inertia, respectively, which are *absolute geometrical properties* of the particles' worldlines. So

> *acceleration is absolute not because a particle accelerates with respect to some absolute space, but because its worldline is curved,*

which is a geometrical fact that is frame-independent (and indeed there is neither motion nor a distinguished space in spacetime). In the same way, inertial motion is absolute – in any reference frame the worldline of an inertial particle is straight. This deep understanding of inertial and accelerated motion in terms of the *shape* of particles' worldlines and *with no reference to space* nicely explains the apparent paradox that seems to have tormented Newton the most – both an inertial and an accelerating particle (appear to) move *in* space, but only the accelerating particle resists its motion. Below we will see that this nice explanation becomes beautiful when it is taken into account that the geometry of a real four-dimensional world is physical geometry which involves real physical objects – worldlines or rather worldtubes in the case of spatially extended bodies.

In general relativity (in curved spacetime) the absoluteness of inertial motion reflects the absolute (frame-independent) geometrical property of the worldline of a free particle to be geodesic. By analogy with the absoluteness of acceleration in flat spacetime, the absoluteness of acceleration in curved spacetime manifest itself in the fact that the worldline of a particle, whose curved-spacetime acceleration ($a^{\mu} = \mathrm{d}^2x^{\mu}/\mathrm{d}\tau^2 + \Gamma^{\mu}_{\alpha\beta}(\mathrm{d}x^{\alpha}/\mathrm{d}\tau)(\mathrm{d}x^{\beta}/\mathrm{d}\tau)$) is different from zero, is not geodesic – the worldline of such a particle is curved or, perhaps more precisely, deformed (intuitively, that deformation can be regarded as an additional curvature to the natural curvature of a geodesic worldline which is due to the curvature of spacetime itself; rigorously, in general relativity a geodesic is not curved, only nongeodesic worldlines are curved). There is a second acceleration in general relativity caused by geodesic deviation, which reflects two facts – that there are no straight worldlines and no parallel or rather congruent worldlines in curved spacetime. This acceleration is not absolute, but relative since it involves two geodesic worldlines (which are not deformed), whereas absolute acceleration involves a single nongeodesic worldline (which is deformed).

Regarding four-dimensional physics as spacetime geometry easily refutes Mach's view of the relativ-

ity of acceleration (now the overwhelming majority of physicists regard acceleration as absolute; here is an example [8.13, p. 34]: *an observer's acceleration is an absolute, local quantity, measurable without reference to anything external*). Two consequences of this view discussed by *Mach* himself [8.14] are i) the equivalence of rotation and translation and ii) the relativity of rotation, which implied (as Mach stated) the equivalence of the Ptolemaic and the Copernican models of our planetary system. It is clear that rotation and translation are distinct in spacetime – the worldline of a particle moving translationally is either a straight line in flat spacetime (when the particle moves uniformly) or a curved line (when the particle accelerates translationally), whereas the worldline of a rotating particle is a helix. This also explains why the Ptolemaic and the Copernican systems are not equivalent – the planets' worldlines are helixes around the worldline of the Sun. Another consequence of Mach's view is that if there were no other bodies in the Universe, one could not talk about the state of motion of a *single* particle since on Mach's view only motion *relative* to a body makes sense. In spacetime the situation is crystal clear – a single particle in the Universe is either a geodesic worldline (which means that the particle moves by inertia) or a deformed worldline (which means that the particle accelerates).

8.2.3 Inertia as Another Manifestation of the Reality of Spacetime

Had Minkowski lived longer he would have certainly noticed that his explanation of the absoluteness of accelerated and inertial motion in terms of the absolute geometrical properties of particles' worldlines (curvature and straightness, respectively) not only reflected the experimental (and therefore absolute, i. e., frame-independent) facts that an accelerating body resists its acceleration, whereas a particle moving by inertia offers no resistance to its motion, but could also *explain* these facts.

Two pieces of reliable knowledge about an accelerating body would have appeared naturally linked in the spacetime explanation of the absoluteness of acceleration – an accelerating body i) resists its acceleration, and ii) is represented by a curved (and therefore *deformed*) worldtube. Then taking into account the reality of the body's worldtube (relativistic length contraction would be impossible if the worldtube of the contracting body were not real as seen from Minkowski's explanation discussed above), would have led to the logically evident, but totally unexpected consequence of linking the two features of the accelerating body – the resistance an accelerating body offers to its acceleration could be viewed as originating from a four-dimensional stress in the deformed worldtube of the body. And it turns out that the *static* restoring force existing in the deformed worldtube of an accelerating body does have the form of the inertial force with which the body resists its acceleration [8.7, Chap. 9]. The origin of the static restoring force (i. e., the inertial force) can be traced down to the most fundamental constituents of matter – as an elementary particle is not a worldline in spacetime its inertia appears to originate from the *distorted* fields which mediate the particle's interactions (the distortion of the fields is caused by the particle's acceleration) [8.7, Chap. 9].

I guess, Minkowski would have been truly thrilled – inertia appears to be another manifestation of the four-dimensionality of the world (since only a *real* worldtube could resist its deformation) along with the other manifestations he knew then – length contraction and all experiments demonstrating that absolute uniform motion could not be detected (that is, that rest and uniform motion could not be distinguished experimentally).

With this insight into the origin of inertia implied by the reality of spacetime, the experimental distinction between accelerated and inertial motion finds a natural but counter-intuitive explanation – an accelerating body resists its acceleration since its worldtube is *deformed* and the static restoring force existing in the worldtube is interpreted as the inertial force, whereas a particle moving by inertia offers no resistance to its motion since its worldtube is *not deformed* – it is straight in flat spacetime and geodesic in curved spacetime – and therefore no restoring force exists in the particle's worldtube (which explains why inertial motion cannot be detected experimentally).

8.2.4 Why Is the Inertial Force Equivalent to the Force of Weight?

The equivalence of the inertial force with which a particle resists its acceleration and the particle's weight (or the gravitational force acting on the particle in terms of the Newtonian gravitational theory) is best visualized by Einstein's thought experiments involving an accelerating elevator and an elevator on the Earth's surface. Assume that a particle is on the floor of an elevator whose acceleration a is equal to the acceleration due

to gravity g. The particle exerts on the elevator's floor an inertial force with which it resists its acceleration (forced on it by the floor). When the same elevator with the same particle on its floor is on the Earth's surface, the particle exerts on the floor a force of the same magnitude which is called the particle's weight (in the Newtonian gravitational theory this force is regarded as the gravitational force acting on the particle). Einstein regarded the equivalence of the two forces as a manifestation of his principle of equivalence according to which the effects of accelerated motion and gravity cannot be distinguished *locally* in spacetime (i. e., for *small* distances and *short* periods of time). In other words, if an observer in a small elevator i) measures the weight of a particle and ii) studies for a short period of time its fall toward the floor of the elevator, he will be unable to determine from his measurements whether the elevator is accelerating with $a = g$ or it is on the Earth's surface.

Initially, Einstein *postulated* the equivalence of the inertial and gravitational forces as part of the principle of equivalence which was a crucial step in the creation of general relativity. Later, when Minkowski's representation of inertial and accelerated motion in spacetime was generalized for the case of curved spacetime it became possible to reveal the deep meaning of this equivalence and of the principle of equivalence itself – inertial and gravitational forces (and masses as will be discussed below) are equivalent since they both are inertial forces (and masses).

By the geodesic hypothesis in general relativity (confirmed by the experimental fact that falling bodies do not resist their fall), a particle falling toward the Earth's surface moves by inertia since its worldtube is geodesic (more precisely, the center of the particle's mass is a geodesic worldline). This means that the particle does not resist its motion in agreement with the fact that its worldtube is not deformed (since it is geodesic).

When the particle reaches the ground it is prevented from moving by inertia (i. e., prevented from falling) and the particle resists the change in its inertial motion. In other words, the particle on the ground is *accelerating* since it is forced by the Earth's surface to *change* its motion by inertia. This counter-intuitive fact – that a particle on the ground accelerates, whereas it is obviously at rest there – is naturally explained by the generalization of Minkowski's observation (see also [8.13, Chap. 9]) that in spacetime an accelerating particle is a curved (deformed) worldtube. Indeed in general relativity the acceleration of a particle at rest on the Earth's surface is the (first) curvature of the particle's worldtube [8.11, pp. 138, 177]. The worldtube of the falling particle is geodesic, but starting at the event at which the particle touches the ground, the particle's worldtube is *constantly* deformed by the huge worldtube of the Earth, which means that the particle on the ground is constantly accelerating (the particle's absolute acceleration is a manifestation of its deformed worldtube).

As the worldtube of the particle, when it is at rest on the ground, is deformed the static restoring force in the worldtube acts back on the Earth's worldtube. This restoring force manifest itself as the resistance force which the particle exerts on the ground, i. e., as the inertial force with which the particle resists its acceleration while being at rest on the ground. Therefore, it becomes clear that what has been traditionally called the gravitational force acting on the particle, or the particle's weight, is in reality the particle's inertial force with which the particle resists its acceleration when it is at rest on the ground. This explains naturally why *there is no such thing as the force of gravity* in general relativity [8.11, p. 109].

To summarize, general relativity showed that what has been traditionally called the force of weight of a particle (or the gravitational force acting on a particle) is the inertial force with which the particle resists its acceleration while being at rest on the Earth's surface. As *Rindler* put it [8.12]:

> *ironically, instead of explaining inertial forces as gravitational ... in the spirit of Mach, Einstein explained gravitational forces as inertial.*

Indeed, according to Mach the origin of inertia is nonlocal since he believed that all the masses in the Universe are responsible for the inertial forces (which implies that these forces are gravitational), whereas the now accepted Minkowski's treatment of acceleration in spacetime (as the curvature of an accelerating particle's worldline) implies that inertia is a *local* phenomenon in spacetime since it originates from the *deformation* of an accelerating particle's worldtube. Therefore inertia is not a nonlocal phenomenon that is caused by the distant masses as Mach argued. One might say that what determines the shape of a free particle's geodesic worldtube (which, when deformed, resists its deformation) are all the masses in the Universe in line with Mach's view. However, that would be misleading since in curved spacetime it is the nearby mass that is essentially responsible for the shape of the geodesics in its vicinity.

The shape of the geodesic worldline of a particle falling toward the Earth, for example, is predominantly determined by the Earth's mass and the distant masses have practically no contribution.

8.2.5 Why Is the Inertial Mass Equivalent to the Gravitational Mass?

When a particle accelerates, the coefficient of proportionality m_i linking the force and the induced by it acceleration in the equation $\boldsymbol{F} = m_i\boldsymbol{a}$ is called the inertial mass of the particle. Since Newton it has been defined as *the measure of the resistance a particle offers to its acceleration.* In the Newtonian gravitational theory, when the same particle is at rest on the Earth's surface the coefficient of proportionality m_g linking the force of gravity and the induced by it acceleration in the equation $\boldsymbol{F} = m_g\boldsymbol{g}$ is called the (passive) gravitational mass of the particle. Since Newton it has been known that the inertial mass and the gravitational mass are equivalent. But no one knew what this equivalence meant. Einstein merely *postulated* it as another manifestation of the principle of equivalence when he created general relativity.

As we saw above the generalization of Minkowski's representation of accelerated and inertial motion for curved spacetime and taking seriously the reality of particles' worldtubes (and the reality of spacetime itself) naturally explained the equivalence of inertial and (what was called before general relativity) gravitational forces. This effectively also explained the equivalence of inertial and gravitational masses – both masses are inertial. Indeed, whether a particle is accelerating or is on the Earth's surface, in both cases the particle is subject to absolute acceleration (since its worldtube is deformed, i. e., nongeodesic) and the particle resists the change in its inertial motion (i. e., resists the deformation of its worldtube). As the inertial mass is the measure of the resistance a particle offers to its acceleration, it does follow that in both cases the particle's mass is inertial.

Since there have been some recent attempts to deny the reality of the relativistic increase of the mass I think it is appropriate to note that those attempts somehow fail to see the obvious reason for the introduction of relativistic mass – *as inertial mass is the measure of the resistance a body offers to its acceleration and as its acceleration is different in different inertial reference frames, the body's inertial mass cannot be the same in all frames* (for more details see [8.7, pp. 114–116]).

8.2.6 Are Gravitational Phenomena Caused by Gravitational Interaction According to General Relativity?

What follows in this section may seem quite controversial but I think it is worth exploring the implications of general relativity *itself* since the generalization of Minkowski's representation of inertial motion for curved spacetime – the geodesic hypothesis – implies that *gravitational phenomena are not caused by gravitational interaction.* Such a stunning possibility [8.15] deserves very serious scrutiny because of its implications for fundamental physics as a whole, and particularly for two research programs as mentioned above – detection of gravitational waves and quantum gravity.

As too much is at stake in terms of both the number of physicists working on quantum gravity and on detection of gravitational waves, and the funds being invested in these worldwide efforts, even the heretical option of not taking gravity for granted should be thoroughly analyzed. It should be specifically stressed, however, that such an analysis may require extra effort from relativists who sometimes appear to be more accustomed to solving technical problems than to examining the physical foundation of general relativity which may involve no calculations. Such an analysis is well worth the effort since it ensures that what is calculated is indeed in the proper framework of general relativity and is not smuggled into it to twist it until it yields some features that resemble gravitational interaction.

Had Minkowski lived longer he would have probably been enormously excited to see his profound idea that four-dimensional physics is spacetime geometry so powerfully boosted by Einstein's discovery that gravitation is a manifestation of the non-Euclidean geometry of spacetime. Indeed, the fact that the appearance of gravitational attraction between two free particles arises from the convergence of their geodesic worldlines in curved spacetime is fully in line with *Minkowski*'s anticipation [8.3, p. 112] that *the laws of physics can find their most complete expression as interrelations between these worldline.* However, keeping in mind how critically and creatively he examined the facts that led to the creation of Einstein's special relativity and how he gave its now accepted spacetime formulation, it is quite reasonable to imagine that Minkowski might have acted in the same way with respect to Einstein's general relativity as well. Imagining such a scenario could help us to examine the logical structure of general relativity by applying the lessons learned from Minkowski's exam-

ination of special relativity. Such an examination now seems more than timely especially in light of the fact that the different approaches aimed at creating a theory of quantum gravity [8.16–18] have been unsuccessful.

In order to explore rigorously the implications of general relativity *itself* let me state explicitly the following facts from it:

- Like flat (Minkowski) spacetime, the non-Euclidean spacetime of general relativity is a *static* entity with a forever given network of worldtubes of macroscopic bodies. Relativists are of course aware of this intrinsic feature of spacetime (reflecting its very nature) – that one cannot talk about dynamics in spacetime [8.13, p. 7]:

  There is no dynamics in spacetime: nothing ever happens there. Spacetime is an unchanging, once-and-for-all picture encompassing past, present, and future.

  But it seems it is not always easy to regard this counter-intuitive feature of spacetime as adequately representing the world.
- The geometry of spacetime is either intrinsic (pseudo-Euclidean in the case of Minkowski spacetime and pseudo-non-Euclidean in the case of de Sitter's vacuum solution of the Einstein–Hilbert equation) or induced by matter (although it is widely assumed to be clear in general relativity that matter causes the curvature of spacetime, that issue is more subtle than usually presented in the literature as briefly discussed below).
- What is still (misleadingly) called the gravitational field in general relativity is not a physical field; at best, the gravitational field can be regarded as a geometrical field.
- There is no gravitational force in general relativity.
- By the geodesic hypothesis, a timelike geodesic in spacetime represents a free partricle, which moves by inertia.

A close examination of these facts reveals that when general relativity is taken for what it is, it does imply that gravitational phenomena are *fully* explained in the theory without the need to assume that they are caused by gravitational interaction. What has the appearance of gravitational attraction between particles involves only *inertial* (*interaction-free*) motion of *free* particles and is merely a result of the curvature of spacetime. In general relativity falling bodies and the planets are all free bodies which move by inertia and for this reason they do not interact in any way with the Earth and the Sun, respectively, since by its very nature *inertial motion does not involve any interaction.*

I think the major reason for so far missing the opportunity to decode *everything* that general relativity has been telling us about the world is that the existence of gravitational interaction has been taken for granted. As a result of adopting such a fundamental assumption without any critical examination, gravitational interaction has been artificially and forcefully inserted into general relativity through i) the definition of a free particle (which posits that otherwise free particles are still subject to gravitational interaction), and ii) the quantity gravitational energy and momentum, which general relativity itself refuses to accommodate.

The often openly stated definition of a free particle in general relativity – a particle is *free from any influences other than the curvature of spacetime* [8.19] – effectively *postulates* the existence of gravitational interaction by almost explicitly asserting that the influence of the spacetime curvature on the *shape* of a free particle's worldline constitutes gravitational interaction.

To see whether a free particle is subject to gravitational interaction, imagine a wandering planet far away from any galaxy which means that in a huge spacetime region the geometry is close to flat and only the planet's mass induces an observable curvature. Imagine also a free particle in that spacetime region, which travels toward the planet. When far away from the planet, the particle's worldline is straight. But as the particle approaches the planet, its worldline becomes increasingly deviated from its straight shape. Despite that its shape changes, the particle's worldline remains geodesic (not deformed) since the curvature of the worldline is simply caused by the spacetime curvature induced by the planet's mass. The standard interpretation of this situation in general relativity, implied by the definition of a free particle, is that the planet, through the spacetime curvature created by its mass, affects the worldline of the particle which is interpreted as gravitational interaction.

However, if carefully analyzed, the assumption that the planet's mass curves spacetime, which in turn changes the shape of the geodesic worldline of a free particle, does not imply that the planet and the particle interact gravitationally. There are four reasons for that.

First, it is assumed that in general relativity the Einstein–Hilbert equation clearly demonstrates that matter determines the geometry of spacetime through the stress energy of matter T_{ab}. In fact, how that happens (how matter curves spacetime) is the major open

question in general relativity. What further complicates the (often taken as self-evident) assertion that matter determines the geometry of spacetime is the fact that in general relativity matter cannot be clearly regarded as something that tells the spacetime geometry how to change since matter itself cannot be defined without that same geometry [8.13, p. 83] –

> T_{ab} *itself is a quantity which refers, not only to* matter, *but also to* geometry

(since T_{ab} contains the metric tensor). Therefore, as very little is known of how matter influences the geometry of spacetime, it is unjustified to take for certain that the change of the shape of the worldline of a free particle by the spacetime curvature caused by a massive body constitutes gravitational interaction; moreover, as indicated below the massive body does not spend any additional energy to change the shape of the particle's worldline.

Second, the shape of the geodesic worldlines of free particles is *naturally* determined by the curvature of spacetime which itself may not be necessarily induced by some mass. This is best seen from the fact that general relativity shows *both* that spacetime is curved by the presence of matter, and that a matter-free spacetime can be *intrinsically* curved. The latter option follows from *de Sitter*'s solution [8.20] of the Einstein–Hilbert equation. Two *test* particles in the de Sitter universe only appear to interact gravitationally since in fact their interaction-like behavior is caused by the *curvature* of their geodesic worldlines (*curvature* here means nonstraightness), which is determined by the *intrinsic* curvature of the de Sitter spacetime. The fact that there are no straight geodesic worldlines in non-Euclidean spacetime (which gives rise to geodesic deviation) manifests itself in the relative acceleration of the test particles toward each other which creates the impression that the particles interact gravitationally. Due to the usual assumption that the masses of *test* particles are negligible in order not to affect the geometry of spacetime, the example with the test particles in the de Sitter universe is a good approximation of a matter-free universe.

Third, the experimental fact that particles of different masses fall toward the Earth with the *same* acceleration in full agreement with general relativity's *a geodesic is particle independent* [8.10, p. 178], ultimately means that the shape of the geodesic worldline of a free particle in spacetime curved by the presence of matter is determined by the spacetime geometry *alone* and not by the matter. This is best seen from the Einstein–Hilbert equation itself – a body curves *solely* spacetime *irrespective* of whether or not there are other particles there, which means that *no additional energy is spent* for *curving* (not deforming) the geodesic worldlines of any free particles that are in the vicinity of the body. That is why a geodesic is particle independent. This feature of general relativity taken alone demonstrates that the fact that the shape of the geodesic worldline of a free particle is determined by the curvature of spacetime does not constitute gravitational interaction.

Fourth, if determining the shape of a free particle's geodesic worldline by the spacetime curvature induced by a body's mass–energy constituted gravitational interaction, that would imply some exchange of *gravitational* energy-momentum between the body and the particle. But there is no such a thing as gravitational energy-momentum in general relativity *itself* – its mathematical structure does not allow a proper tensorial expression for a gravitational energy-momentum. This counter-intuitive feature of general relativity is not surprising at all since i) there is no *physical* gravitational field (one can use the term *field* to describe gravitational phenomena only in the sense of a *geometrical* field, but such a field describes the geometry of spacetime and as such does not possess any energy), and ii) there is no gravitational force and therefore there is no gravitational energy either since such energy is defined as the work done by gravitational forces.

In short, the mass-energy of a body influences the geometry of spacetime no matter whether or not there are any particles in the body's vicinity, and the shape of a free particle's geodesic worldline reflects the spacetime curvature no matter whether it is intrinsic or induced by matter.

8.2.7 Is There Gravitational Energy?

Although this question was answered above it is necessary to explain briefly why the energy involved in gravitational phenomena is not gravitational. Consider the energy of oceanic tides which is transformed into electrical energy in tidal power stations. The tidal energy is part of gravitational phenomena, but is not gravitational energy. It seems most appropriate to call it *inertial energy* because it originates from the work done by inertial forces acting on the blades of the tidal turbines – the blades further deviate the volumes of water from following their geodesic (inertial) paths (the water volumes are already deviated since they are prevented from falling) and the water volumes *resist* the

further change in their inertial motion; that is, the water volumes exert *inertial* forces on the blades. With respect to the resistance, this example is equivalent to the situation in hydroelectric power plants where water falls on the turbine blades from a height (this example is even clearer) – the blades prevent the water from falling (i. e., from moving by inertia) and it resists that change. It is that resistance force (i. e., inertial force) that moves the turbine, which converts the inertial energy of the falling water into electrical energy. According to the standard explanation it is the kinetic energy of the falling water (originating from its potential energy) that is converted into electrical energy. However, it is evident that behind the kinetic energy of the moving water is its inertia (its resistance to its being prevented from falling) – it is the inertial force with which the water acts on the turbine blades when prevented from falling. And it can be immediately seen that the inertial energy of the falling water (the work done by the inertial force on the turbine blades) is equal to its kinetic energy [8.15, Appendix B].

8.2.8 Do Gravitational Waves Carry Gravitational Energy?

At present there exists a widespread view that there is indirect astrophysical evidence for the existence of gravitational energy. That evidence is believed to come from the interpretation of the decrease of the orbital period of binary pulsar systems, notably the system PSR 1913+16 discovered by *Hulse* and *Taylor* in 1974 [8.21]; recently it was also reported of *evidence for the loss of orbital energy in agreement ... with the emission of gravitational waves* from a binary system of two candidate black holes [8.22, 23]. According to this interpretation the decrease of the orbital period of such binary systems is caused by the loss of energy due to gravitational waves emitted by the systems. Almost without being challenged (with only few exceptions [8.24–26]) this view holds that quadrupole radiation of gravitational waves which carry gravitational energy away from the binary systems has been indirectly experimentally confirmed.

I think the interpretation that the orbital motion of the neutron stars in the PSR 1913+16 system, for example, loses energy by emission of gravitational waves should be rigorously reexamined since it *contradicts general relativity*, particularly the geodesic hypothesis and the experimental evidence which confirmed it. The reason is that by the geodesic hypothesis the neutron stars, whose worldlines had been regarded as exact *geodesics* (since the stars had been *modeled dynamically as a pair of orbiting point masses* [8.27]), *move by inertia without losing energy* because the very essence of inertial motion is motion without any loss of energy. For this reason no energy can be carried away by the gravitational waves emitted by the binary pulsar system. Therefore, the experimental fact of the decay of the orbital motion of PSR 1913+16 (the shrinking of the stars' orbits) cannot be regarded as evidence for the existence of gravitational energy. The observed diminishing of the orbital period of the binary pulsar should be caused by other mechanisms, e.g., magnetic or (and) tidal effects. Tidal friction was suggested in 1976 [8.28] as an alternative to the explanation given by Hulse and Taylor, which ignored the tidal effects by treating the neutron stars as point masses. The argument that the neutron stars would behave as rigid bodies (since they are believed to be very compact) is not convincing because by the same reason – the large spacetime curvature caused by the stars (which is ultimately responsible for their rigidity) – the other gravitational effects, that is, the tidal effects, are also very strong.

If it really turns out that binary pulsars are not slowed by the emission of gravitational energy (as I believe it would), that would be another important lesson of the superior role of physics over mathematics in physical theories. Being aware that not devoting particular attention to physical (conceptual) analyses of physical situations could lead to problems. *Wheeler* and *Taylor* stressed that the superiority of physics should always and explicitly be kept in mind in what he called the first moral principle [8.29]:

> *Never make a calculation until you know the answer. Make an estimate before every calculation, try a simple physical argument (symmetry! invariance! conservation!) before every derivation, guess the answer to every paradox and puzzle.*

In the case of the decrease of the orbital period of binary systems, the physical argument is that the geodesic hypothesis and the statement that bodies, whose worldlines are *geodesic*, emit gravitational energy cannot be both correct. Another physical argument in the case of binary systems involves orbital energy. Saying that a binary system of two neutron stars has orbital (gravitational) energy is equivalent to saying that two bodies in uniform relative motion approach each other in flat spacetime also have some common energy since in both cases only inertial motion is involved – the stars' worldlines are geodesic in curved spacetime and the bodies worldlines are straight in flat spacetime. The two cases

are equivalent since the stars also move by inertia and there is no exchange of some gravitational energy between them – as discussed above the same stress-energy tensor of each star produces the same spacetime curvature no matter whether or not the other star is there.

8.2.9 Can Gravity Be Quantized?

In the case of physical interactions when one talks about the energy associated with an interaction, it is the energy of the entity (the field and its quanta) that mediates an interaction and it is that entity and its energy which are quantized. What should make us to consider seriously the possibility that a theory of quantum gravity might be impossible is the fact that there is no such a thing as an entity which mediates gravitational interaction in general relativity. Although the term *gravitational field* is widely used in the general relativistic literature its correct meaning is to describe the geometry of spacetime and nothing more. It is not a physical field that can be quantized. If the gravitational field represented some physical entity, it should be measurable. *Misner*, *Thorne*, and *Wheeler* paid special attention to the question of the measurement of the gravitational field [8.30, p. 399]:

> I know how to measure the electromagnetic field using test charges; what is the analogous procedure for measuring the gravitational field? *This question has, at the same time, many answers and none.*
>
> *It has no answers because nowhere has a precise definition of the term* gravitational field *been given – nor will one be given. Many different mathematical entities are associated with gravitation: the metric, the Riemann curvature tensor, the Ricci curvature tensor, the curvature scalar, the covariant derivative, the connection coefficients, etc. Each of these plays an important role in gravitation theory, and none is so much more central than the others that it deserves the name* gravitational field. *Thus it is that throughout this book the terms* gravitational field *and* gravity *refer in a vague, collective sort of way to all of these entities. Another, equivalent term used for them is the* geometry of spacetime.

As there is no physical entity which is represented by the term *gravitational field* in general relativity it does follow that there is no energy and momentum of that nonexistent physical entity. This in turn should make us to accept the unambiguous fact that the logical structure of general relativity does not contain and does not allow a *tensor* of the gravitational energy and momentum. It was Einstein who first tried to insert the concept of gravitational energy and momentum *forcefully* into general relativity (since he represented it by a pseudo-tensor, not a tensor as it should be) in order to ensure that gravity can still be regarded as some interaction. Einstein made the gigantic step toward the profound understanding of gravity as spacetime curvature but even he seems to have been unable to accept all implications of the revolutionary view of gravitational phenomena.

For decades the efforts of many brilliant physicists to create a quantum theory of gravity have not been successful. This could be an indication that those efforts might not have been in the right direction. In such desperate times in fundamental physics all approaches and ideas should be on the research table, including the approach discussed here – that general relativity completely explains gravitational phenomena without the need of gravitational interaction, if gravity is consistently and rigorously regarded as a manifestation of the non-Euclidean geometry of spacetime, that is, general relativity implies that gravitational phenomena are not caused by gravitational interaction. An immediate implication of this approach is that quantum gravity understood as quantization of gravitational interaction is impossible because there is nothing to quantize. If this turns out to be the case, the efforts to quantize the apparent gravitational interaction should be redirected toward what seems to be the actual open question in gravitational physics – how matter curves spacetime – since it is quantum physics which should deal with this question and which should provide the definite answer to the central question in general relativity – whether or not there exists some kind of interaction between physical bodies mediated by spacetime itself.

8.3 Propagation of Light in Noninertial Reference Frames in Spacetime

So far the issue of the propagation of light in noninertial reference frames (accelerating in special relativity and associated with a body in general relativity whose worldline is not geodesic) has not been fully presented in the books on relativity (I am aware only of two books where only the slowing down of light in curved spacetime is explicitly discussed [8.31, 32]). This issue has a straightforward and self-evident explanation when the physical phenomenon of propagation of light is regarded as spacetime geometry. In fact, regarding the phenomenon of light propagation as spacetime geometry naturally explains both why the speed of light is the same in all inertial reference frames in flat spacetime and why it is not constant in noninertial reference frames (in flat and curved spacetimes).

Let us start with the propagation of light in inertial reference frames. In his 1905 paper Einstein merely postulated (as his second postulate) that the speed of light is the same in all inertial frames. It is clear now that Einstein did not need to introduce a second postulate in special relativity since the constancy of the speed of light follows from the first postulate (the principle of relativity) – the consequence of Maxwell's equations that electromagnetic waves propagate with a constant speed (which turned out to be a fundamental constant $c = (\varepsilon_0\mu_0)^{-1/2}$) should hold in all inertial frames. However, at the time when Michelson and Morley proved experimentally that the speed of light is constant and a bit later when Einstein postulated it, that fact had been a complete mystery.

The situation completely changed in 1908 when Minkowski gave the four-dimensional formulation of special relativity. One of the implications of Minkowski's four-dimensional physics was the explanation of the constancy of the speed of light in inertial frames. In the ordinary *three-dimensional* (space *and* time) language, the speed of light is the same in all inertial frames because each reference frame has not only its own (proper) time, but also (as Minkowski showed) its own (proper) space and light propagates with respect to the proper space of each frame and the frame's proper time measures the duration of the light propagation.

However, complete understanding of the whole phenomenon of propagation of light is obtained when the physics of this phenomenon is regarded as spacetime geometry. Only when Minkowski gave the spacetime formulation of special relativity it was revealed that there are three kinds of length in spacetme and that the propagation of light is represented by null (or lightlike) geodesics whose status is absolute or frame independent. A light signal which travels the distance $\mathrm{d}x$ for the time period $\mathrm{d}t$ in *any* inertial reference frame is represented in flat spacetime by a lightlike worldline whose length between the events of emission and arrival of the light signal is zero (in the case of a two-dimensional spacetime)

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}x^2 = 0\,.$$

It is evident from here that in any inertial reference frame the speed of light is the same: $c = \mathrm{d}x/\mathrm{d}t$.

However, even in flat spacetime the spacetime metric in a noninertial reference frame (e.g., an elevator accelerating with a proper acceleration a along the x-axis) is [8.30, p. 173]

$$\mathrm{d}s^2 = \left(1 + \frac{ax}{c^2}\right)^2 c^2\,\mathrm{d}t^2 - \mathrm{d}x^2\,. \tag{8.1}$$

It is immediately seen from here that for a lightlike worldline (representing a propagating light signal) $\mathrm{d}s^2 = 0$ and therefore the coordinate anisotropic velocity of light c^{a} at a point x is

$$c^{\mathrm{a}}(x) = \pm c\left(1 + \frac{ax}{c^2}\right), \tag{8.2}$$

where the $+$ and $-$ signs correspond to the propagation of a light signal along or against the x-axis, respectively.

As spacetime is flat it is clear that the nonconstancy of the velocity of light in an accelerating elevator is not caused by the curvature of spacetime. It is seen from (8.1) that the non-Euclidean metric in the accelerating elevator results from the curvature of the elevator's worldline along which the time axis is constantly chosen (at each point of the elevator's worldline the time axis is the tangent at that point and coincides with the time axis of the instantaneously comoving inertial reference frame at that point). In 1960 *Synge* stressed the need to distinguish between two types of effects in relativity [8.11]:

> *Spacetime is either flat or curved, and in several places in the book I have been at considerable pains to separate truly gravitational effects due to curvature of space-time from those due to curvature of the observer's world-line (in most ordinary cases the latter predominate).*

The anisotropic velocity of light (8.2) is another manifestation of the latter effect.

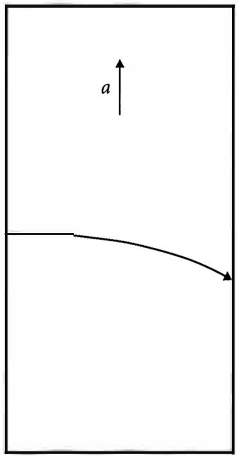


Fig. 8.3 A horizontal light ray propagates in an accelerating elevator

That the velocity of light is not constant in an accelerating elevator was first realized by *Einstein* whose thought experiments involving an accelerating elevator and an elevator at rest on the Earth's surface led him to the discovery that a horizontal light signal bends in such elevators (as shown in Fig. 8.3) [8.33]:

> *A curvature of rays of light can only take place when the velocity of propagation of light varies with position.*

The implications of these results have not been fully explored. Although the bending of a horizontal light ray in Einstein's original thought experiments with elevators found their way even in introductory physics textbooks [8.34–37], the obvious question of whether light rays propagating in a vertical direction (parallel and antiparallel to the elevator's acceleration) are also affected by the elevators's acceleration, has never been asked. The definite answer to this question could have been given even before Minkowski's spacetime formulation of special relativity.

Consider an inertial reference frame I in which an elevator is at rest. At a given moment t_0 the elevator starts to accelerate upward as shown in Fig. 8.4 [8.7, Sect. 7.3]. The x-axes of I and a noninertial frame N associated with the elevator are along the elevator's acceleration. At the same moment t_0 three light rays are emitted *simultaneously* in the elevator from points D, A, and C toward point B. As at that moment I and N are at rest the emission of the light rays is simultaneous in I as well (now we can say that I is the instantaneously comoving inertial frame at the moment t_0 which means that I and N share the same instantaneous space and therefore they share the same class of simultaneous events at t_0).

At the next moment as N accelerates an observer in I sees that the three light rays arrive simultaneously not at point B, but at B′ (since during the time the light rays travel the elevator, i. e., N, moves upward); the inertial observer sees that the horizontal light ray emitted from D propagates along a straight line (the dashed yellow line in Fig. 8.4). Let DB = AB = BC = r in I. Since for the time $t = r/c$ in I the light rays travel toward B, the elevator moves a distance $\delta = at^2/2 = ar^2/2c^2$. As the simultaneous arrival of the three rays at point B′ as viewed in I is an absolute (observer-independent) fact due to its being a *single* event, it follows that the rays arrive simultaneously at B′ as seen by an observer in N as well.

We have DB = AB = BC = r in both I and N because this thought experiment represents a clearly nonrelativistic situation and therefore the relativistic contraction of AB and BC in I can safely be ignored (the elevator just started to accelerate and its velocity relative to I is negligible compared to c). Since for the *same* coordinate time $t = r/c$ in N, the three light rays travel *different* distances DB′ $\approx r$, AB′ $= r + \delta$, and CB′ $= r - \delta$, before arriving *simultaneously* at point B′, an observer in the elevator concludes that the propagation of light is affected by the elevator's acceleration. The *average* velocity $c^{\mathrm{a}}_{\mathrm{AB'}}$ of the light ray propagating from A to B′ is slightly greater than c

$$c^{\mathrm{a}}_{\mathrm{AB'}} = \frac{r+\delta}{t} \approx c\left(1 + \frac{ar}{2c^2}\right).$$

The average velocity $c^{\mathrm{a}}_{\mathrm{B'C}}$ of the light ray propagating from C to B′ is slightly smaller than c

$$c^{\mathrm{a}}_{\mathrm{CB'}} = \frac{r-\delta}{t} \approx c\left(1 - \frac{ar}{2c^2}\right).$$

It is easily seen that to within terms proportional to c^{-2} the average light velocity between A and B is equal to that between A and B′, i. e., $c^{\mathrm{a}}_{\mathrm{AB}} = c^{\mathrm{a}}_{\mathrm{AB'}}$ and also $c^{\mathrm{a}}_{\mathrm{CB}} = c^{\mathrm{a}}_{\mathrm{CB'}}$

$$c^{\mathrm{a}}_{\mathrm{AB}} = \frac{r}{t - \frac{\delta}{c}} = \frac{r}{t - \frac{at^2}{2c}} = \frac{c}{1 - \frac{ar}{2c^2}} \approx c\left(1 + \frac{ar}{2c^2}\right) \tag{8.3}$$

and

$$c^{\mathrm{a}}_{\mathrm{CB}} = \frac{r}{t + \frac{\delta}{c}} \approx c\left(1 - \frac{ar}{2c^2}\right). \tag{8.4}$$

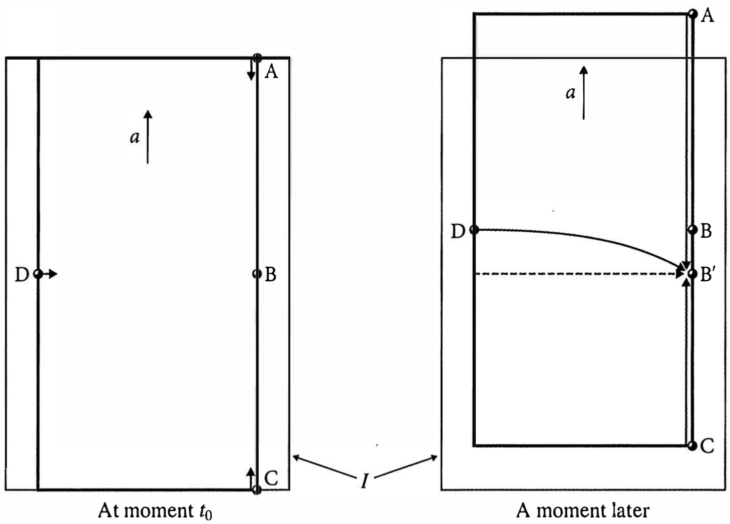


Fig. 8.4 How an inertial observer I and an observer in an accelerating elevator see the propagation of three light rays in the elevator

Since the *coordinate* time t is involved in the calculation of the average velocities (8.3) and (8.4), it is clear that these expressions represent the average *coordinate* velocities between the points A and B and the points C and B, respectively.

The same expressions for the average coordinate velocities $c^{\mathrm{a}}_{\mathrm{AB}}$ and $c^{\mathrm{a}}_{\mathrm{CB}}$ can also be obtained from the expression for the coordinate velocity of light (8.2) in N. As the coordinate velocity $c^{\mathrm{a}}(x)$ is continuous on the interval $[x_{\mathrm{A}}, x_{\mathrm{B}}]$, one can calculate the average coordinate velocity between A and B in Fig. 8.4

$$\begin{aligned} c^{\mathrm{a}}_{\mathrm{AB}} &= \frac{1}{x_{\mathrm{B}} - x_{\mathrm{A}}} \int_{x_{\mathrm{A}}}^{x_{\mathrm{B}}} c^{\mathrm{a}}(x)\,\mathrm{d}x \\ &= c\left(1 + \frac{a x_{\mathrm{B}}}{c^2} + \frac{ar}{2c^2}\right), \end{aligned} \tag{8.5}$$

where we have taken into account the fact that $x_{\mathrm{A}} = x_{\mathrm{B}} + r$. When the coordinate origin is at point B ($x_{\mathrm{B}} = 0$), the expression (8.5) coincides with (8.3). In the same way,

$$c^{\mathrm{a}}_{\mathrm{BC}} = c\left(1 + \frac{a x_{\mathrm{B}}}{c^2} - \frac{ar}{2c^2}\right), \tag{8.6}$$

where $z_{\mathrm{C}} = x_{\mathrm{B}} - r$. For $x_{\mathrm{B}} = 0$, (8.6) coincides with (8.4).

Analogous expressions can be obtained for the average coordinate velocity of light in an elevator at rest on the Earth's surface, which is subject to the acceleration due to gravity g [8.7, Sect. 7.3]

$$c^{\mathrm{g}}_{\mathrm{AB}} = c\left(1 + \frac{g x_{\mathrm{B}}}{c^2} + \frac{gr}{2c^2}\right) \tag{8.7}$$

and

$$c^{\mathrm{g}}_{\mathrm{BC}} = c\left(1 + \frac{g x_{\mathrm{B}}}{c^2} - \frac{gr}{2c^2}\right). \tag{8.8}$$

As indicated above representing the physical situation depicted in Fig. 8.4 in terms of spacetime geometry is the best way to demonstrate that it is the *curvature* of the worldline of point B (and B′) which causes

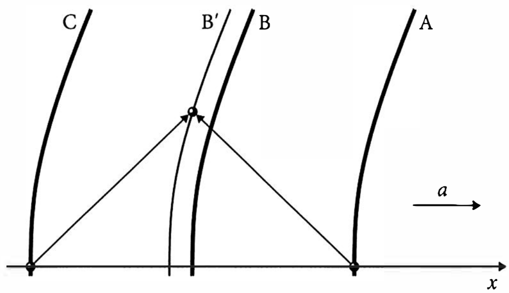


Fig. 8.5 Regarding the physical phenomenon of light propagation as spacetime geometry provides a straightforward explanation of the anizotropic propagation of light in the accelerating elevator – the nonconstancy of the velocity of light observed in the elevator is caused by the curvature of the worldline of point B

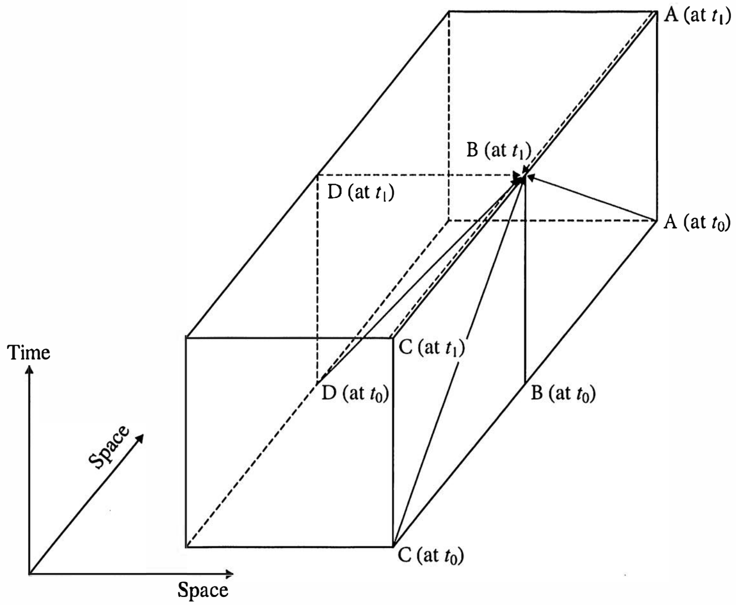


Fig. 8.6 The spacetime geometry of the propagation of three light rays in an inertial elevator

the nonconstancy of the velocity of light in N. This is done in Fig. 8.5 which represents a two-dimensional spacetime diagram. The worldlines of points A, B, B′, and C as well as the worldlines of the light rays emitted from A and C are depicted in the figure. It is obvious that due to the curvature of the worldline of B (and B′) the worldlines of the light rays meet at the worldline of B′, not at the worldline of B. In this thought experiment it is the curvature of the worldline B alone which is responsible for the anisotropic velocity of light in the accelerating elevator, but in more complex experiments with light rays in an accelerating elevator the curvature of the worldlines of the light sources and the light detectors causes the anisotropy in the propagation of light in noninertial reference frames.

The spacetime geometry of the propagation of all three light rays emitted from points A, C, and D can be represented in a three-dimensional spacetime diagram. In order to make the spacetime diagram of the accelerating elevator more easily understandable, let us first consider the propagation of the three light rays in an elevator, which moves with constant velocity as shown in Fig. 8.6. The elevator at the moment t_0, when the three light rays are emitted simultaneously toward point B, is represented by the bottom side of the parallelepiped in Fig. 8.6. At moment t_1, when the worldlines of the three light rays meet at the worldline of point B, the elevator is represented by the top side of the parallelepiped.

Now consider the spacetime diagram in Fig. 8.7 showing the propagation of the three light rays in an accelerating elevator. The elevator at moments t_0 and t_1 is represented by the bottom and top sides of the parallelepiped, respectively (the two sides of the parallelepiped represent the instantaneous spaces of the noninertial reference frame associated with the elevator, which correspond to the moments t_0 and t_1). It is again quite obvious that what causes the anisotropic propagation of light in the accelerating elevator (in this specific thought experiment) is the curvature of the worldline of point B – the worldlines of the light rays emitted from A, C, and D at t_0 all meet at the worldline of point B′.

It turns out that the average coordinate velocity of light is not sufficient for the complete description of propagation of light in noninertial reference frames. The average coordinate velocity of light explains the propagation of light in such frames in situations like the one discussed above. However, in a situation where the average light velocity between two points – a source and an observation point – is determined *with respect to one of the points*, where the local velocity of light is c and where the *proper time* is used, that average velocity of light is not coordinate; it can be regarded as an average *proper* velocity of light. For instance, such a situation occurs in the Shapiro time delay effect.

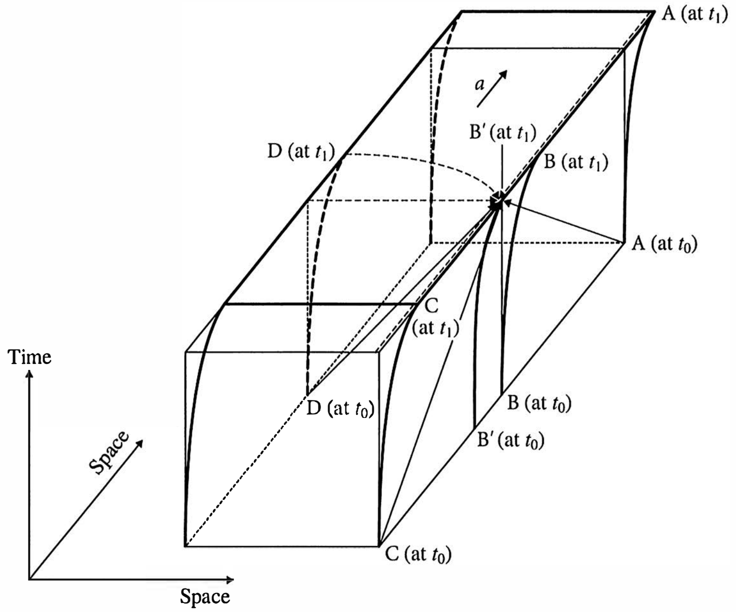


Fig. 8.7 The spacetime geometry of the propagation of three light rays in an accelerating elevator. In order not to complicate the spacetime diagram, the elevator at moments t_0 and t_1 is shown as the *bottom* and *top sides* of the *parallelepiped*. However, in reality the two sides (representing the elevator at t_0 and t_1 and also the instantaneous spaces at t_0 and t_1 of the noninertial reference frame N associated with the elevator) are not parallel, because those sides (i. e., the instantaneous spaces of N at the two moments) coincide with the spaces of the instantaneously comoving inertial reference frames at t_0 and t_1, which are not parallel since the two instantaneously comoving inertial reference frames are in relative motion

We calculated the average coordinate velocity of light in an accelerating elevator, but now we will determine the average proper velocity of light in a noninertial reference frame N associated with an elevator at rest on the Earth's surface (Fig. 8.8). The reason is to explain in detail how light propagates toward and away from the Earth since this issue is not always explained properly in introductory physics textbooks. For example, one can read that *a beam of light will accelerate in a gravitational field, just like objects that have mass* and therefore *near the surface of the earth, light will fall with an acceleration of* 9.81 m/s^2 [8.34]. We shall now see that during its *fall* toward the Earth, light is slowing down – a negative acceleration of 9.81 m/s^2 is decreasing its velocity.

As an elevator at rest on the Earth is prevented from falling it is accelerating (since its worldtube is curved) with an acceleration g due to gravity.

To calculate the average proper velocity of light which originates from B and is observed at A, we have to determine the initial velocity of a light signal at B and its final velocity at A, both with respect to A [8.7, Sect. 7.4]. As the local velocity of light is c, the final velocity of the light signal determined at A is obviously c. By taking into account that in a parallel *gravitational field*, proper and coordinate distances are the same [8.38], we can determine the initial velocity of the light signal at B as seen from A

$$c_{\mathrm{B}}^{\mathrm{g}} = \frac{\mathrm{d}x_{\mathrm{B}}}{\mathrm{d}\tau_{\mathrm{A}}} = \frac{\mathrm{d}x_{\mathrm{B}}}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}\tau_{\mathrm{A}}} .$$

Here $\mathrm{d}\tau_{\mathrm{A}} = \mathrm{d}s_{\mathrm{A}}/c$ is the proper time of an observer with constant spatial coordinates at A,

$$\mathrm{d}\tau_{\mathrm{A}} = \left(1 + \frac{gx_{\mathrm{A}}}{c^2}\right)\mathrm{d}t ,$$

and $\mathrm{d}x_{\mathrm{B}}/\,\mathrm{d}t = c^{\mathrm{g}}(x_{\mathrm{B}})$ is the coordinate velocity of light at B,

$$c^{\mathrm{g}}(x_{\mathrm{B}}) = c\left(1 + \frac{gx_{\mathrm{B}}}{c^2}\right) ,$$

which follows from the metric (the line element) in the case of parallel *gravitational field* [8.30, p. 1056]

$$\mathrm{d}s^2 = \left(1 + \frac{gx}{c^2}\right)^2 c^2\,\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 .$$

As $x_{\mathrm{A}} = x_{\mathrm{B}} + r$ (we again have $\mathrm{AB} = \mathrm{BC} = r$) and $gx_{\mathrm{A}}/c^2 < 1$ (since for any value of x in N, there exists the restriction $|x| < c^2/g$), for the coordinate time $\mathrm{d}t$, we have (to within terms $\propto c^{-2}$)

$$\mathrm{d}t \approx \left(1 - \frac{gx_{\mathrm{A}}}{c^2}\right)\mathrm{d}\tau_{\mathrm{A}} = \left(1 - \frac{gx_{\mathrm{B}}}{c^2} - \frac{gr}{c^2}\right)\mathrm{d}\tau_{\mathrm{A}} .$$

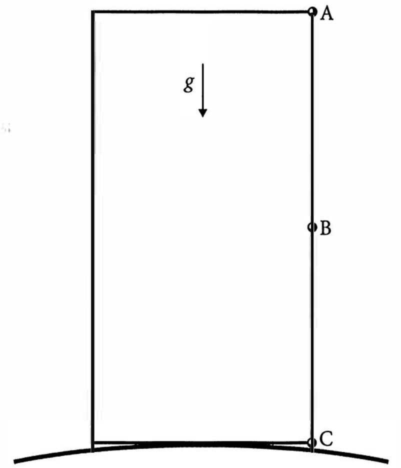


Fig. 8.8 An elevator at rest on the Earth's surface

Then for the initial velocity c_{B}^{g} at B as determined at A, we obtain

$$c_{\mathrm{B}}^{g} = c\left(1 + \frac{gx_{\mathrm{B}}}{c^2}\right)\left(1 - \frac{gx_{\mathrm{B}}}{c^2} - \frac{gr}{c^2}\right),$$

or, keeping only the terms proportional to c^{-2},

$$c_{\mathrm{B}}^{g} = c\left(1 - \frac{gr}{c^2}\right). \tag{8.9}$$

Therefore, an observer at A will determine that when a light signal is emitted at B with the initial velocity (8.9) during the time of its journey toward A (away from the Earth's surface) it will *accelerate* with an acceleration g and will arrive at A with a final velocity equal to c.

For the average proper velocity $\bar{c}_{\mathrm{BA}}^{g} = (1/2)(c_{\mathrm{B}}^{g} + c)$ of light propagating from B to A as observed at A, we have

$$\bar{c}_{\mathrm{BA}}^{g} \quad \text{(as observed at A)} = c\left(1 - \frac{gr}{2c^2}\right). \tag{8.10}$$

As the local velocity of light at A (measured at A) is c, it follows that if a light signal propagates from A toward B, its initial velocity at A is c, the final velocity of the light signal at B is (8.9) and therefore, as seen from A, it is subject to a negative acceleration g and will *slow down* as it *falls* toward the Earth. The average proper velocity $\bar{c}_{\mathrm{AB}}^{g}$ (as seen from A) of a light signal emitted at A with the initial velocity c and arriving at B with the final velocity (8.9) will be equal to the average proper velocity $\bar{c}_{\mathrm{BA}}^{g}$ (as seen from A) of a light signal propagating from B toward A. Thus, as seen from A, the back and forth average proper speeds of light travelling between A and B are the *same*.

Now let us determine the average proper velocity of light between B and A with respect to point B. A light signal emitted at B as seen from B will have an initial (local) velocity c there. The final velocity of the signal at A as seen from B will be

$$c_{\mathrm{A}}^{g} = \frac{\mathrm{d}x_{\mathrm{A}}}{\mathrm{d}\tau_{\mathrm{B}}} = \frac{\mathrm{d}x_{\mathrm{A}}}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}\tau_{\mathrm{B}}},$$

where $\mathrm{d}x_{\mathrm{A}}/\,\mathrm{d}t = c^{g}(x_{\mathrm{A}})$ is the coordinate velocity of light at A,

$$c^{g}(x_{\mathrm{A}}) = c\left(1 + \frac{gx_{\mathrm{A}}}{c^2}\right),$$

and $\mathrm{d}\tau_{\mathrm{B}}$ is the proper time at B,

$$\mathrm{d}\tau_{\mathrm{B}} = \left(1 + \frac{gx_{\mathrm{B}}}{c^2}\right)\mathrm{d}t\,.$$

Then as $x_{\mathrm{A}} = x_{\mathrm{B}} + r$, we obtain for the velocity of light at A, as determined at B,

$$c_{\mathrm{A}}^{g} = c\left(1 + \frac{gr}{c^2}\right). \tag{8.11}$$

Using (8.11), the average proper velocity of light propagating from B to A as determined from B becomes

$$\bar{c}_{\mathrm{BA}}^{g} \quad \text{(as observed at B)} = c\left(1 + \frac{gr}{2c^2}\right). \tag{8.12}$$

If a light signal propagates from A to B, its average proper velocity $\bar{c}_{\mathrm{AB}}^{g}$ (as seen from B) will be equal to $\bar{c}_{\mathrm{BA}}^{g}$ (as seen from B) – the average proper speed of light propagating from B to A. This demonstrates that, for an observer at B, a light signal emitted from B with velocity c will *accelerate* toward A with an acceleration g and will arrive there with the final velocity (8.11). As determined by the B-observer, a light signal emitted from A with initial velocity (8.11) will be *slowing down* (with $-g$) as it *falls* toward the Earth and will arrive at B with a final velocity equal to c. Therefore, an observer at B will agree with an observer at A that a light signal will *accelerate* with an acceleration g on its way from B to A and will *decelerate* while *falling* toward the Earth during its propagation from A to B, but will disagree on the velocity of light at the points A and B.

The use of the average anisotropic velocity of light in the Shapiro time delay and the Sagnac effect is demonstrated in [8.7, Sects. 7.5, 7.8].

The calculation of the average proper velocity of light in an accelerating frame is obtain in the same way

and gives [8.7, Sect. 7.4]

$$\bar{c}^{\mathrm{g}}_{\mathrm{BA}} \quad \text{(as observed at A)} = c\left(1 - \frac{ar}{2c^2}\right) \qquad (8.13)$$

and

$$\bar{c}^{\mathrm{a}}_{\mathrm{BA}} \quad \text{(as observed at B)} = c\left(1 + \frac{ar}{2c^2}\right), \qquad (8.14)$$

where a is the proper acceleration of the frame.

Comparing the average coordinate velocities of light (8.5) and (8.6) with (8.7) and (8.8) and the average proper velocities of light (8.13) and (8.14) with (8.10) and (8.12) shows that their expressions are the same in an accelerating elevator and in an elevator on the Earth's surface. This fact can be regarded as another manifestation of the equivalence principle. But this principle only *postulates* such equivalences without any explanation; they are pure mystery. The complete explanation of the identical anisotropy in the propagation of light in both elevators is obtained only when the phenomenon of propagation of light is regarded as geometry of a real spacetime. Only then it becomes clear that acceleration is a curvature of a worldline. Only then it becomes clear that, like an accelerating elevator, an elevator on the Earth's surface also accelerates since its worldtube, like the worldtube of the accelerating elevator, is also *curved*. Then the same accelerations $a = g$ of the elevators demonstrate that their worldtubes are equally curved, which causes the identical anisotropic propagation of light in an accelerating elevator and in an elevator at rest on the Earth's surface. The fact that the worldlines of the points of the accelerating elevator are as much deviated from their geodesic shapes (i. e., from their straight shapes in flat spacetime) as the worldlines of the points of the elevator on the Earth's surface are deviated from their geodesic shapes in curved spacetime naturally explains the equivalence of all physical phenomena in the elevators (which equivalence was postulated as the equivalence principle).

References

8.1 M. Planck: *Eight Lectures On Theoretical Physics Delivered at Columbia University in 1909* (Columbia Univ. Press, New York 1915) pp. 129–130, translated by A. P. Wills

8.2 N.D. Mermin: What's bad about this habit?, Phys. Today **62**, 8 (2009)

8.3 H. Minkowski: Space and time. In: *Space and Time: Minkowski's Papers on Relativity*, ed. by V. Petkov (Minkowski Institute Press, Montreal 2012)

8.4 H. Minkowski: The relativity principle. In: *Space and Time: Minkowski's Papers on Relativity*, ed. by V. Petkov (Minkowski Institute Press, Montreal 2012) pp. 42–43

8.5 H.A. Lorentz: *The Theory of Electrons and Its Applications to the Phenomena of Light and Radiant Heat*, 2nd edn. (Dover, Mineola New York 2003), see also his comment on p. 321

8.6 H. Minkowski: *Space and Time: Minkowski's Papers on Relativity* (Minkowski Institute Press, Montreal 2012)

8.7 V. Petkov: *Relativity and the Nature of Spacetime*, 2nd edn. (Springer, Heidelberg 2009)

8.8 G.F.R. Ellis, R.M. Williams: *Flat and Curved Space-Times* (Oxford Univ. Press, Oxford 1988)

8.9 D.F. Wallace: *Everything and More: A Compact History of Infinity* (Norton, New York 2003) p. 259

8.10 W. Rindler: *Relativity: Special, General, and Cosmological* (Oxford Univ. Press, Oxford 2001) p. 178

8.11 J.L. Synge: *Relativity: The General Theory* (North-Holland, Amsterdam 1960), p. ix

8.12 W. Rindler: *Essential Relativity*, 2nd edn. (Springer, New York 1977) p. 244

8.13 R. Geroch: *General Relativity: 1972 Lecture Notes* (Minkowski Institute Press, Montreal 2013)

8.14 E. Mach: *The Science of Mechanics*, 6th edn. (La Salle, Illinois 1960)

8.15 V. Petkov: *Inertia and Gravitation: From Aristotle's Natural Motion to Geodesic Worldlines in Curved Spacetime* (Minkowski Institute Press, Montreal 2012), Chap. 6 and Appendix C

8.16 J. Murugan, A. Weltman, G.F.R. Ellis (Eds.): *Foundations of Space and Time: Reflections on Quantum Gravity* (Cambridge Univ. Press, Cambridge 2011)

8.17 B. Booß-Bavnbek, G. Esposito, M. Lesch (Eds.): *New Paths Towards Quantum Gravity* (Springer, Berlin Heidelberg 2010)

8.18 D. Oriti (Ed.): *Approaches to Quantum Gravity: Toward a New Understanding of Space, Time and Matter* (Cambridge Univ. Press, Cambridge 2009)

8.19 J.B. Hartle: *Gravity: An Introduction to Einstein's General Relativity* (Addison Wesley, San Francisco 2003) p. 169

8.20 W. de Sitter: Over de relativiteit der traagheid: Beschouingen naar aanleiding van Einstein's hypothese, K. Akad. Wet. Amst. **25**, 1268–1276 (1917)

8.21 R.A.J.H.T. Hulse: Discovery of a pulsar in a binary system, Astrophys. J. **195**, L51–L53 (1975)

8.22 M.J. Valtonen, H.J. Lehto: Outbursts in OJ287: A new test for the general theory of relativity, Astrophys. J. **481**, L5–L7 (1997)

8.23 M.J. Valtonen, H.J. Lehto, K. Nilsson, J. Heidt, L.O. Takalo, A. Sillanpää, C. Villforth, M. Kidger, G. Poyner, T. Pursimo, S. Zola, J.-H. Wu, X. Zhou, K. Sadakane, M. Drozdz, D. Koziel, D. Marchev, W. Ogloza, C. Porowski, M. Siwak, G. Stachowski, M. Winiarski, V.-P. Hentunen, M. Nissinen, A. Liakos, S. Dogru: A massive binary black-hole system in OJ 287 and a test of general relativity, Nature **452**, 851–853 (2008)
8.24 N. Rosen: Does gravitational radiation exist?, Gen. Relativ. Gravit. **10**, 351–364 (1979)
8.25 F.I. Cooperstock: Does a dynamical system lose energy by emitting gravitational waves?, Mod. Phys. Lett. **A14**, 1531 (1999)
8.26 F.I. Cooperstock: The role of energy and a new approach to gravitational waves in general relativity, Ann. Phys. **282**, 115–137 (2000)
8.27 J.H. Taylor, J.M. Weisberg: Further experimental tests of relativistic gravity using the binary pulsar PSR 1913+16, Astrophys. J. **345**, 434–450 (1989)
8.28 S.A. 8albus, K. 8recher: Tidal friction in the binary pulsar system PSR 1913+16, Astrophys. J. **203**, 202–205 (1976)
8.29 E.F. Taylor, J.A. Wheeler: *Spacetime Physics*, 2nd edn. (Freeman, New York 1992) p. 20
8.30 C.W. Misner, K.S. Thorne, J.A. Wheeler: *Gravitation* (Freeman, San Francisco 1973)
8.31 H. Ohanian, R. Ruffini: *Gravitation and Spacetime*, 2nd edn. (W.W. Norton, New York 1994), Sect. 4.4
8.32 E.F. Taylor, J.A. Wheeler: *Exploring Black Holes: Introduction to General Relativity* (Addison Wesley Longman, San Francisco 2000), pp. 5–3, E-1
8.33 A. Einstein: *Relativity: The Special and General Theory*, The Masterpiece Science Edition (Pi Press, New York 2005) p. 97
8.34 P.A. Tipler: *Physics*, Vol. 3, 4th edn. (Freeman, New York 1999) p. 1272
8.35 R.L. Reese: *University Physics*, Vol. 2 (Brooks/Cole, New York 2000) p. 1191
8.36 R.A. Serway: *Physics*, Vol. 2, 4th edn. (Saunders, Chicago 1996) p. 1180
8.37 P.M. Fishbane, S. Gasiorowicz, S.T. Thornton: *Physics* (Prentice Hall, New Jersey 1993) p. 1192
8.38 W. Rindler: Counterexample to the Lenz-Schiff argument, Am. J. Phys. **36**, 540 (1968)

9. Electrodynamics of Radiating Charges in a Gravitational Field

Øyvind Grøn

The electrodynamics of a radiating charge and its electromagnetic field based upon the Lorentz–Abraham–Dirac (LAD) equation are discussed both with reference to an inertial reference frame and a uniformly accelerated reference frame. It is demonstrated that energy and momentum are conserved during runaway motion of a radiating charge and during free fall of a charge in a field of gravity. This does not mean that runaway motion is really happening. It may be an unphysical solution of the LAD equation of motion of a radiating charge due to the unrealistic point particle model of the charge upon which it is based. However it demonstrates the consistency of classical electrodynamics, including the LAD equation which is deduced from Maxwell's equations and the principle of energy-momentum conservation applied to a radiating charge and its electromagnetic field. The decisive role of the Schott energy in this connection is made clear and an answer is given to the question: What sort of energy is the Schott energy and where is it found? It is the part of the electromagnetic field energy which is proportional to (minus) the scalar product of the velocity and acceleration of a moving accelerated charged particle. In the case of the electromagnetic field of a point charge it is localized at the particle. This energy is negative if the acceleration is in the same direction as the velocity and positive if it is in the opposite direction. During runaway motion the Schott energy becomes more and more negative and in the case of a charged particle with finite extension, it is localized in a region with increasing extension surrounding the particle. The Schott energy provides the radiated energy of a freely falling charge. Also it is pointed out that a proton and a neutron fall with the same acceleration in a uniform gravitational field, although the proton radiates and the neutron does not. It is made clear that the question as to whether or not a charge radiates has a reference-dependent answer. An accelerated charge is not observed to radiate by an observer comoving with the charge, although an inertial observer finds that it radiates.

9.1 The Dynamics of a Charged Particle

The analysis of the energy-momentum balance of a radiating charge is usually based on the equation of motion of a point charge. The nonrelativistic version of the equation was discussed already more than 100 years ago by *H.A. Lorentz* [9.1]. The relativistic generalization of the equation was originally found by

M. Abraham [9.2] in 1905 and re-derived in 1909 by *M. von Laue* [9.3] who Lorentz transformed the nonrelativistic equation from the instantaneous rest frame of the charge to an arbitrary inertial frame. A new deduction of the Lorentz covariant equation of motion was given by *P. A. M. Dirac* in 1938 [9.4]. This equation is therefore called the Lorentz–Abraham–Dirac equation, or for short the LAD equation. A particularly interesting feature about Dirac's deduction is that it establishes a connection between Maxwell's equations and the equation of motion for a charged particle. It shows that the presence of the Abraham four-vector in the equation of motion (Sect. 9.1.2) comes from conservation of energy and momentum for a closed system consisting of a charge and its electromagnetic field.

9.1.1 The Nonrelativistic Equation of Motion of a Radiating Charge

In the nonrelativistic limit the equation of motion of a radiating charge, q, with mass m_0, acted upon by an external force, f_{ext}, takes the form

$$m_0\ddot{\boldsymbol{r}} = \boldsymbol{f}_{\text{ext}} + m_0\tau_0\dddot{\boldsymbol{r}}\,, \qquad \tau_0 = \frac{q^2}{6}\pi\varepsilon_0 m_0 c^3\,, \tag{9.1}$$

where the dot denotes differentiation with respect to the (Newtonian) time. If q and m_0 represent the charge and mass of an electron, respectively, τ_0 is of the same order of magnitude as the time taken by light to move a distance equal to the classical electron radius, i. e., $\tau_0 \approx 10^{-23}$ s. The general solution of the equation is

$$\ddot{\boldsymbol{r}}(T) = \mathrm{e}^{T/\tau_0} \times \left[\ddot{\boldsymbol{r}}(0) - \frac{1}{m\tau_0}\int_0^T \mathrm{e}^{-T'/\tau_0}\boldsymbol{f}_{\text{ext}}(T')\,\mathrm{d}T'\right]. \tag{9.2}$$

Hence the charge performs a runaway motion unless one chooses the initial condition

$$m\tau_0\ddot{\boldsymbol{r}}(0) = \int_0^\infty \mathrm{e}^{-T'/\tau_0}\boldsymbol{f}_{\text{ext}}(T')\,\mathrm{d}T'\,. \tag{9.3}$$

By combining (9.2) and (9.3) one obtains [9.5]

$$m\ddot{\boldsymbol{r}}(T) = \int_0^\infty \mathrm{e}^{-s}\boldsymbol{f}_{\text{ext}}(T + \tau_0 s)\,\mathrm{d}s\,. \tag{9.4}$$

This equation shows that the acceleration of the charge at a point of time T is determined by the future force, weighted by a decreasing exponential factor with value 1 at the time T, and a time constant τ_0, i. e., there is *pre-acceleration.*

In his discussion of (9.1) *Lorentz* [9.1] writes:

In many cases the new force represented by the second term in (9.1) may be termed a resistance *to the motion. This is seen if we calculate the work of the force during an interval of time extending from $T = T_1$ to $T = T_2$. The result is*

$$\int_{T_1}^{T_2} \dot{\boldsymbol{a}}\cdot\boldsymbol{v}\,\mathrm{d}T = [\boldsymbol{a}\cdot\boldsymbol{v}]_{T_1}^{T_2} - \int_{T_1}^{T_2} \boldsymbol{a}^2\,\mathrm{d}T\,. \tag{9.5}$$

Here the first term disappears if, in the case of periodic motion, the integration is extended to a full period, and also if at the instants T_1 and T_2 either the velocity or the acceleration is zero. Whenever the above formula reduces to the last term, the work of the force is seen to be negative, so that the name of resistance is then justly applied.

P. Yi [9.6] gives the following interpretation:

The total energy of the system may be split into three pieces: the kinetic energy of the charged particle, the radiation energy, and the electromagnetic energy of the Coulomb field. In effect, the last acts as a sort of energy reservoir that mediates the energy transfer from the first to the second and in the special case of uniform acceleration provides all the radiation energy without extracting any from the charged particle.

9.1.2 The Relativistic Equation of Motion of a Radiating Charge

The original relativistic equation of motion of a particle with rest mass m_0 and charge q (the LAD equation) may be written as [9.7]

$$F^{\mu}_{\text{ext}} + \Gamma^{\mu} = m_0\dot{U}^{\mu}\,, \tag{9.6}$$

where

$$\Gamma^{\mu} \equiv m_0\tau_0\left(\dot{A}^{\mu} - A^{\alpha}A_{\alpha}U^{\mu}\right), \tag{9.7}$$

and the dot denotes differentiation with respect to the proper time of the particle. Here F^{μ}_{ext} is the external force acting upon the particle, U^{μ} is its four-velocity

and A^μ its four-acceleration. (Capital letters shall be used for four-vector components referring to an inertial frame, and units are used so that $c = 1$.)

The vector Γ^μ is called the *Abraham four-force* and is given by

$$\Gamma^\mu = \gamma\left(\boldsymbol{v}\cdot\boldsymbol{\Gamma}, \boldsymbol{\Gamma}\right), \tag{9.8}$$

where $\boldsymbol{\Gamma}$ is the three-dimensional force called the *field reaction force* [9.8], v is the ordinary velocity of the particle, and $\gamma = (1-v^2)^{-1/2}$. In an inertial reference frame the Abraham four-force may be written as

$$\Gamma^\mu = m_0\tau_0\gamma(\boldsymbol{v}\cdot\dot{\boldsymbol{g}}, \dot{\boldsymbol{g}}), \tag{9.9}$$

where $g = (A_\alpha A^\alpha)^{\frac{1}{2}}$ is the proper acceleration of the charged particle in the inertial frame. Hence

$$\boldsymbol{\Gamma} = m_0\tau_0\dot{\boldsymbol{g}}. \tag{9.10}$$

In flat spacetime there exist global inertial frames. However, in curved spacetime there are only *local* inertial frames. They are freely falling. Then $\boldsymbol{g}$ is the acceleration of the particle in a freely falling frame in which the particle is instantaneously at rest. In such a frame a freely falling particle has no acceleration. Hence, a particle falling freely in a gravitational field has vanishing four-acceleration. From (9.7) and (9.10) is seen that for such a particle the Abraham four-force vanishes. This is also valid for a charged particle emitting radiation while it falls. This case shall be treated in more detail in Sect. 9.5.

According to the Lorentz covariant Larmor formula, valid with reference to inertial systems, the energy radiated by the particle per unit time is (using the sign convention that the signature of the metric is +2),

$$P_{\mathrm{L}} = m_0\tau_0 A^\alpha A_\alpha = m_0\tau_0 g^2. \tag{9.11}$$

The radiated momentum per unit proper time is

$$P_{\mathrm{R}}^\mu = P_{\mathrm{L}} U^\mu. \tag{9.12}$$

From the equation of motion (9.6) we obtain the energy equation

$$\begin{aligned}\boldsymbol{v}\cdot\boldsymbol{F}_{\mathrm{ext}} &= \gamma^{-1}\left(m_0\dot{U}^0 - \Gamma^0\right)\\ &= m_0\gamma^3\boldsymbol{v}\cdot\boldsymbol{a} - \boldsymbol{v}\cdot\boldsymbol{\Gamma}\\ &= \frac{\mathrm{d}E_{\mathrm{K}}}{\mathrm{d}T} - \boldsymbol{v}\cdot\boldsymbol{\Gamma},\end{aligned} \tag{9.13}$$

where $E_{\mathrm{K}} = (\gamma-1)m_0c^2$ is the kinetic energy of the particle and T is the coordinate time in the inertial frame. Note that the energy supplied by the external force is equal to the change of the kinetic energy of the charge when the Abraham four-force vanishes. Hence, it is tempting to conclude from the Abraham–Lorentz theory, i. e., from (9.10) and (9.13), that a charge having constant acceleration does not radiate. This is, however, not the case. The power due to the field reaction force is

$$\begin{aligned}\boldsymbol{v}\cdot\boldsymbol{\Gamma} &= m_0\tau_0\frac{\mathrm{d}}{\mathrm{d}T}(\gamma^4\boldsymbol{v}\cdot\boldsymbol{a}) - P_{\mathrm{L}}\\ &= -\frac{\mathrm{d}E_{\mathrm{S}}}{\mathrm{d}T} - \frac{\mathrm{d}E_{\mathrm{R}}}{\mathrm{d}T},\end{aligned} \tag{9.14}$$

where E_{R} is the energy of the radiation field and E_{S} is the Schott energy defined by

$$E_{\mathrm{S}} \equiv -m_0\tau_0\gamma^4\boldsymbol{v}\cdot\boldsymbol{a} = -m_0\tau_0 A^0. \tag{9.15}$$

(This energy was called *acceleration energy* by *Schott* [9.9] but is now usually called *Schott energy*.) Hence, in the case of constant acceleration, when the Abraham four-force vanishes, the charge radiates in accordance with Larmor's formula, (9.11), and the rate of radiated energy is equal to minus the rate of change of the Schott energy. The energy equation may now be written as

$$\frac{\mathrm{d}W_{\mathrm{ext}}}{\mathrm{d}T} = \boldsymbol{v}\cdot\boldsymbol{F}_{\mathrm{ext}} = \frac{\mathrm{d}}{\mathrm{d}T}(R_{\mathrm{K}} + E_{\mathrm{S}} + E_{\mathrm{R}}), \tag{9.16}$$

where W_{ext} is the work on the particle due to the external force.

Let $\boldsymbol{P}_{\mathrm{ext}}$ be the momentum delivered to the particle from the external force. Then $\mathrm{d}\boldsymbol{P}_{\mathrm{ext}}/\mathrm{d}T = \boldsymbol{F}_{\mathrm{ext}}$, and by means of (9.1), (9.2), and (9.7) we obtain

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{P}_{\mathrm{ext}}}{\mathrm{d}T} &= \boldsymbol{F}_{\mathrm{ext}} = m_0\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}T} - m_0\tau_0\left(\frac{\mathrm{d}\boldsymbol{A}}{\mathrm{d}T} - g^2\boldsymbol{v}\right)\\ &= \frac{\mathrm{d}\boldsymbol{P}_M}{\mathrm{d}T} + \frac{\mathrm{d}\boldsymbol{P}_{\mathrm{S}}}{\mathrm{d}T} + \frac{\mathrm{d}\boldsymbol{P}_{\mathrm{R}}}{\mathrm{d}T}.\end{aligned} \tag{9.17}$$

Thus, according to (9.12) and (9.17) the four-momentum of the particle takes the form

$$P^\mu = P_M^\mu + P_{\mathrm{S}}^\mu, \tag{9.18}$$

where

$$P_M^\mu = m_0 U^\mu, \quad P_{\mathrm{S}}^\mu = -m_0\tau_0 A^\mu \tag{9.19}$$

are the mechanical four-momentum of the particle and the Schott four-momentum, respectively. In addition we have the four-momentum of the radiation field, which is not a state function of the particle.

9.1.3 Significance of the Schott Momentum

In order to demonstrate as clearly as possible the necessity of taking into account the Schott momentum in the dynamics of a charged particle we shall here consider circular motion with a constant speed [9.9]. Then $\boldsymbol{v} \cdot \boldsymbol{a} = 0$ and the Schott energy vanishes, but not the Schott momentum. It is

$$\boldsymbol{p}_\mathrm{S} = -m_0 \tau_0 \gamma^2 \boldsymbol{a} \,. \tag{9.20}$$

Since both the kinetic energy and the Schott energy are constant, (9.16) in this case reduces to

$$\frac{\mathrm{d}W_\mathrm{ext}}{\mathrm{d}T} = \boldsymbol{v} \cdot \boldsymbol{F}_\mathrm{ext} = \frac{\mathrm{d}E_\mathrm{R}}{\mathrm{d}T} = P_\mathrm{L} \,. \tag{9.21}$$

This equation shows that the radiated energy is provided by the tangential component of the external force.

Although the radiated energy per unit time is equal to the power due to the tangential component of the external force, the radiated momentum is not due only to this force. In order to see this most clearly we insert the expression for the centripetal acceleration into the expression (9.20) for the Schott momentum, which gives

$$\boldsymbol{p}_\mathrm{S} = -m_0 \tau_0 \gamma^2 \frac{v^2}{r} \boldsymbol{e}_r \,, \tag{9.22}$$

where $\boldsymbol{e}_r$ is the radial unit vector. The rate of change of the Schott momentum with respect to the inertial laboratory time is

$$\frac{\mathrm{d}\boldsymbol{p}_\mathrm{S}}{\mathrm{d}T} = m_0 \tau_0 \gamma^2 \frac{v^2}{r^2} \boldsymbol{v} \,. \tag{9.23}$$

Putting $\boldsymbol{F} = \boldsymbol{F}_\parallel + \boldsymbol{F}_\perp$ where $\boldsymbol{F}_\parallel$ and $\boldsymbol{F}_\perp$ are the components of $\boldsymbol{F}$ along and orthogonal to $\boldsymbol{v}$, we obtain

$$\boldsymbol{F}_\perp = \gamma m_0 \boldsymbol{a} \,, \tag{9.24}$$

and

$$\boldsymbol{F}_\parallel = \frac{\mathrm{d}\boldsymbol{p}_\mathrm{S}}{\mathrm{d}T} + P_\mathrm{L} \boldsymbol{v} \,. \tag{9.25}$$

For an uncharged particle the centripetal force $\boldsymbol{F}_\perp$ is the only force. But in the case of a charged particle a tangential force $\boldsymbol{F}_\parallel$ is necessary to keep the velocity constant. The radiated energy comes from the work performed by this force. The radiated momentum is partly due to $\boldsymbol{F}_\parallel$ and partly due to the change of the direction of the Schott momentum vector.

9.2 Schott Energy as Electromagnetic Field Energy

Already in 1915 *Schott* [9.10] argued that in the case of uniformly accelerated motion

> *the energy radiated by the electron is derived entirely from its acceleration energy; there is as it were internal compensation amongst the different parts of its radiation pressure, which causes its resultant effect to vanish.*

But what is the *acceleration energy*, now called *the Schott energy*? *Schott* [9.10] and later *Rohrlich* [9.8] noted that there is an important difference between the radiation rate and the rate of change of the Schott energy:

> *The radiation rate is always positive (or zero) and describes an* irreversible loss *of energy; the Schott energy changes in a* reversible *fashion, returning to the same value whenever the state of motion repeats itself.*

Rohrlich also wrote [9.11]:

> *If the Schott energy is expressed by the electromagnetic field, it would describe an energy content of the near field of the charged particle which can be changed* reversibly. *In periodic motion energy is borrowed, returned, and stored in the near-field during each period. Since the time of energy measurement is usually large compared to such a period only the average energy is of interest and that average of the Schott energy rate vanishes. Uniformly accelerated motion permits one to borrow energy from the near-field for large* macroscopic *time-intervals, and no averaging can be done because at no two points during the motion is the acceleration four-vector the same. Nobody has so far shown in detail just how the Schott energy occurs in the near-field, how it is stored, borrowed etc.*

A step toward answering this challenge was taken by *C. Teitelboim* [9.12]. He made a Lorentz invariant separation of the field tensor of the electromagnetic field of a point charge into two parts, $F^{\mu\nu} = F_{\text{I}}^{\mu\nu} + F_{\text{II}}^{\mu\nu}$, where $F_{\text{I}}^{\mu\nu}$ is the velocity field and $F_{\text{II}}^{\mu\nu}$ the acceleration field. Inserting these parts into the expression for the energy-momentum tensor of the electromagnetic field, Teitelboim found that the energy-momentum tensor contains terms of three types: a part $T_{\text{I,I}}^{\mu\nu}$ independent of the acceleration, a part $T_{\text{I,II}}^{\mu\nu}$ depending linearly upon the acceleration, and a part $T_{\text{II,II}}^{\mu\nu}$ depending linearly upon the square of the acceleration of the charged particle producing the fields. Teitelboim then defined $T_{\text{I}}^{\mu\nu} = T_{\text{I,I}}^{\mu\nu} + T_{\text{I,II}}^{\mu\nu}$ and $T_{\text{II}}^{\mu\nu} = T_{\text{II,II}}^{\mu\nu}$. The contribution of the interference between the fields I and II has been included in $T_{\text{I}}^{\mu\nu}$, whereas the tensor $T_{\text{II}}^{\mu\nu}$ is related only to the part of the field depending upon the square of the acceleration. Teitelboim showed that the energy-momentum associated with the field $F_{\text{II}}^{\mu\nu}$ travels with the speed of light. The field fronts are spheres with centers at the emission points. The four-momentum associated with $T_{\text{I}}^{\mu\nu}$ remains bound to the charge. Furthermore he calculated the four-momenta and their time derivatives associated with $T_{\text{I}}^{\mu\nu}$ and $T_{\text{II}}^{\mu\nu}$.

The results of Rohrlich and Teitelboim have been summarized by *P. Pearle* [9.13] in the following way:

> *The term Γ^{μ} in the Lorentz–Dirac equation, as given in (9.6), is called the Abraham force. Its first term, $m_0\tau_0\dot{A}^{\mu}$ is called the Schott term, and its second, $-m_0\tau_0 A^{\alpha}A_{\alpha}U^{\mu}$, the radiation reaction term. The zeroth component of the radiation reaction term is to be interpreted as the radiation rate. Indeed, the scalar product of this term with U_{μ} is the relativistic version of the Larmor formula. The spatial component of this term, proportional to $-\mathbf{v}$ like a viscous drag force, may similarly be interpreted as the radiation reaction force of the electron.*
>
> *The physical meaning of the Schott term has been puzzled over for a long time. Its zero component represents a power which adds* Schott acceleration energy *to the electron and its associated electromagnetic field. The work done by an external force not only goes into electromagnetic radiation and into increasing the electron's kinetic energy, but it causes an increase in the* Schott acceleration energy *as well. This change can be ascribed to a change in the* bound *electromagnetic energy in the electron's induction field, just as the last term of (9.14) can be ascribed to a change in the* free *electromagnetic energy in the electron's radiation field.*
>
> *What meaning should be given to the Schott term?* Teitelboim [9.12] *has argued convincingly that when an electron accelerates, its near-field is modified so that a correct integration of the electromagnetic four-momentum of the electron includes not only the Coulomb four-momentum $(q^2/8\pi\varepsilon_0 r)U^{\mu}$, but an extra four-momentum $-m_0\tau_0 A^{\mu}$ of the bound electromagnetic field.*

It remained to obtain a more precise localization of the Schott field energy.

Rowe [9.15] modified Teitelboim's separation of the energy-momentum tensor of the electromagnetic field of a point charge, which he described by a delta-function, and introduced a separation into three symmetrical, divergence-free parts. In order to obtain a finite expression for the localization of the Schott energy as part of the energy of the electromagnetic field of a charged particle, *Eriksen* and *Grøn* [9.7] applied Rowe's separation to a charged particle with a finite radius and obtained the following result. The Schott energy is inside a spherical light front S touching the front end of a moving Lorentz contracted charged particle.

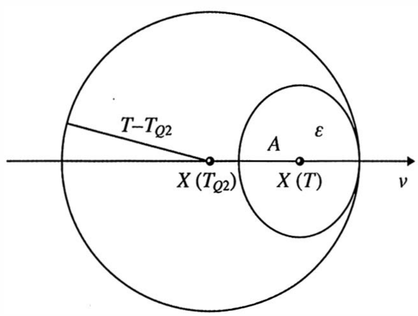


Fig. 9.1 A Lorentz contracted charged particle with proper radius ε moving to the right with velocity v. The field is observed at a point of time T, and at this moment the center of the particle is at the position $X(T)$. The circle is a field front produced at the retarded point of time T_{Q2} when the center of the particle was at the position $X(T_{Q2})$. The field front is chosen such that it just touches the front of the particle. The Schott energy is localized in the shaded region between the field front and the ellipsoid representing the surface of the particle. The velocity is chosen to be $v = 0.6$ (after [9.14], courtesy of the American Association of Physics Teachers)

From Fig. 9.1 one finds that at the point of time T the radius of the light front S that represents the boundary of the distribution of the Schott energy, is

$$T - T_{Q2} = \varepsilon\sqrt{\frac{1+v}{1-v}}\,, \tag{9.26}$$

where the field at the light front S is produced at the retarded point of time T_{Q2}, ε is the proper radius of the particle, and v is the absolute value of its velocity. Hence, unless the velocity of the charge is close to that of light, the Schott energy is localized just outside the surface of the charged particle. The radius of the light front S increases towards infinity for the field of a charge approaching the velocity of light, for example during runaway motion, which we shall now consider in Sect. 9.4.

In the deduction of the localization of the Schott energy we have not applied the equation of motion of the charge. We have only considered the field produced by the charge. Hence the deduction permits us to consider a charge with a finite proper radius. However, the LAD equation is deduced for a point charge. So relating the Schott energy to the conservation of energy for a radiating point charge and the field it produces, we should take the limit $\varepsilon \to 0$. In this limit it seems that the Schott energy is localized at the point charge. But it must be admitted that this limit seems rather unphysical, and our conclusion should rather be that this limit signals a breakdown of classical electrodynamics, or at least some sort of unsolved problem.

A slightly different perception of the Schott energy, still interpreted as electromagnetic field energy, has recently been given by *D. R. Rowland* [9.16]. He found that the Schott energy is the difference between the energy in the actual bound electromagnetic field of a charge and the energy in the bound field if the charge had moved with a constant velocity equal to its instantaneous velocity. Rowland's analysis further provides the following physical explanation of the existence of the Schott energy:

> *This difference arises because the bound fields of a charge cannot respond* rigidly *when the state of motion of a charge is changed by an external force. During uniform acceleration, the rate of change of this difference is just the negative of the rate at which radiation energy is created, and hence the power needed to accelerate a charged particle uniformly is just that which is required to accelerate a neutral particle with the same rest mass even though the charge is radiating.*

9.3 Pre-Acceleration and Schott Energy

The LAD equation has two strange consequences [9.17–25]: pre-acceleration, which is accelerated motion before a force acts; and run-away motion, which is accelerated motion of a charge after a force which acted upon it has ceased to act.

It has been claimed that during a period of pre-acceleration, before a charge is acted upon by an external force, the charge will not emit radiation [9.26]. In this section we will review a recent demonstration we have given where it was shown that a charge emits radiation during a period of pre-acceleration and that the radiation energy then comes from the Schott energy, which decreases during this period [9.27].

We shall consider a particle with charge Q and rest mass m_0 moving in an inertial frame and acted upon by an external force F of finite duration.

With P_{R}^{μ} as defined in (9.12) and P_{S}^{μ} in (9.19), the LAD equation can be written as

$$F^{\nu} = m_0\dot{U}^{\nu} + \dot{P}_{\mathrm{S}}^{\nu} + \dot{P}_{\mathrm{R}}^{\nu}\,. \tag{9.27}$$

In the following we restrict ourselves to linear motion (along the x-axis). The equations can be simplified by introducing the rapidity of the particle,

$$\alpha = \operatorname{artanh} v\,. \tag{9.28}$$

Hence,

$$v = \tanh\alpha\,,\quad \gamma = \cosh\alpha\,,\quad \gamma v = \sinh\alpha\,,$$

$$a = \frac{\mathrm{d}v}{\mathrm{d}t} = \frac{\mathrm{d}\tau}{\mathrm{d}t}\frac{\mathrm{d}v}{\mathrm{d}\tau} = \frac{\dot{v}}{\cosh\alpha} = \frac{\dot{\alpha}}{\cosh^3\alpha}\,, \tag{9.29}$$

$$U^{\nu} = (\cosh\alpha, \sinh\alpha, 0, 0)\,,$$
$$A^{\nu} = \dot{U}^{\nu} = (\dot{\alpha}\sinh\alpha, \dot{\alpha}\cosh\alpha, 0, 0)\,, \tag{9.30}$$

where $\dot{\alpha} = a_0$, i. e. $\dot{\alpha}$ is the acceleration in the inertial rest frame. The components of the mechanical, Schott, and radiation four-momenta may then be expressed as

$$m_0U^{\nu} = m_0(\cosh\alpha, \sinh\alpha)\,, \tag{9.31}$$

$$P_S^\nu = -\frac{2}{3}Q^2\dot{\alpha}(\sinh\alpha, \cosh\alpha), \quad (9.32)$$

$$P_R^\nu = \frac{2}{3}Q^2 \int_{-\infty}^{\tau} \dot{\alpha}^2(\cosh\alpha, \sinh\alpha)\,d\tau. \quad (9.33)$$

Differentiation gives

$$m_0\dot{U}^\nu = m_0\dot{\alpha}(\sinh\alpha, \cosh\alpha), \quad (9.34)$$

$$\begin{aligned}\dot{P}_S^\nu &= -\frac{2}{3}Q^2\ddot{\alpha}(\sinh\alpha, \cosh\alpha)\\ &\quad -\frac{2}{3}Q^2\dot{\alpha}^2(\cosh\alpha, \sinh\alpha),\end{aligned} \quad (9.35)$$

$$\dot{P}_R^\nu = \frac{2}{3}Q^2\dot{\alpha}^2(\cosh\alpha, \sinh\alpha) \quad (9.36)$$

with the sum

$$\begin{aligned}m_0\dot{U}^\nu + \dot{P}_S^\nu + \dot{P}_R^\nu &= \left(m_0\dot{\alpha} - \frac{2}{3}Q^2\ddot{\alpha}\right)\\ &\quad \times(\sinh\alpha, \cosh\alpha).\end{aligned} \quad (9.37)$$

In terms of the rapidity the Minkowski force (9.27) reads as

$$F^\nu = (\gamma v F, \gamma F) = F(\sinh\alpha, \cosh\alpha). \quad (9.38)$$

The LAD equation for linear motion then takes the form [9.28]

$$F = m_0(\dot{\alpha} - \tau_0\ddot{\alpha}). \quad (9.39)$$

Note that (9.39) transforms into the nonrelativistic equation of motion when α is replaced by v and proper time by laboratory time [9.29]. Equation (9.39) may be written as

$$\frac{d}{d\tau}\left(e^{-\tau/\tau_0}\dot{\alpha}\right) = -\frac{F}{m_0\tau_0}e^{-\tau/\tau_0}. \quad (9.40)$$

Let τ_1 and τ_2 be two points of proper time with $\tau_1 < \tau_2$ and F a function of τ such that $F(\tau) = 0$ for $\tau < \tau_1$ and $\tau > \tau_2$. Then the general solution of (9.40) may be written as

$$e^{-\tau/\tau_0}\dot{\alpha}(\tau) = \int_{\tau}^{\tau_2} \frac{F(\tau')}{m_0\tau_0}e^{-\tau'/\tau_0}\,d\tau' + C_0, \quad (9.41)$$

where C_0 is a constant. For $\tau > \tau_2$ the integral is zero, and

$$\begin{aligned}&\dot{\alpha}(\tau) = C_0 e^{\tau/\tau_0},\\ \text{i.e.}\quad &\alpha(\tau) = C_0\tau_0 e^{\tau/\tau_0} + \text{const.}\end{aligned} \quad (9.42)$$

When $C_0 \neq 0$ this is a *run away* solution. The rapidity increases without any boundary, and the velocity approaches the velocity of light when $\tau \to \infty$.

In this section we put $C_0 = 0$, which gives the following solution of (9.40),

$$\dot{\alpha}(\tau) = e^{\tau/\tau_0}\int_{\tau}^{\tau_2} \frac{F(\tau')}{m_0\tau_0}e^{-\tau'/\tau_0}\,d\tau'. \quad (9.43)$$

The integral has the same value for all $\tau < \tau_1$ and is equal to zero for $\tau > \tau_2$. For convenience we introduce the notation

$$f(\tau) \equiv \int_{\tau}^{\tau_2} \frac{F(\tau')}{m_0\tau_0}e^{-\tau'/\tau_0}\,d\tau'. \quad (9.44)$$

We put $\alpha(-\infty) = 0$ and obtain from (9.43) and (9.44), for

$$\begin{aligned}&\text{for}\quad \tau < \tau_1, \quad \dot{\alpha} = e^{\tau/\tau_0}f(\tau_1),\\ &\alpha = \tau_0 e^{\tau/\tau_0}f(\tau_1),\end{aligned} \quad (9.45)$$

$$\begin{aligned}&\text{for}\quad \tau_1 < \tau < \tau_2, \quad \dot{\alpha} = e^{\tau/\tau_0}f(\tau),\\ &a = \tau_0 e^{\tau/\tau_0}f(\tau) + \frac{1}{m_0}\int_{\tau_1}^{\tau} F(\tau')\,d\tau',\end{aligned} \quad (9.46)$$

$$\begin{aligned}&\text{for}\quad \tau > \tau_2, \quad \dot{\alpha} = 0,\\ &\alpha = \alpha(\tau_2) = \frac{1}{m_0}\int_{\tau_1}^{\tau_2} F(\tau')\,d\tau'.\end{aligned} \quad (9.47)$$

Note that $\dot{\alpha} = 0$ for $\tau < \tau_1$ if $f(\tau_1) = 0$. That is, there is no pre-acceleration if $\int_{\tau_1}^{\tau_2} F(\tau')e^{-\tau'/\tau_0}\,d\tau' = 0$.

In order to discuss the energy and momentum of the particle and its field, we consider the formulation (9.27) of the LAD equation, which is a conservation equation of energy and momentum in differential form. Let τ_a and τ be two points of proper time with $\tau > \tau_a$. Then according to (9.27)

$$\int_{\tau_a}^{\tau} F^\nu\,d\tau = \Delta\left(m_0 U^\nu\right) + \Delta P_S^\nu + \Delta P_R^\nu. \quad (9.48)$$

For $\nu = 0$ the left-hand side is the work done by the external force and for $\nu = 1$ it is the delivered momentum. The Δ symbols refer to the increments from τ_a to τ.

As seen from (9.45) the energies and momenta in the pre-acceleration period are given by (9.31)–(9.33)

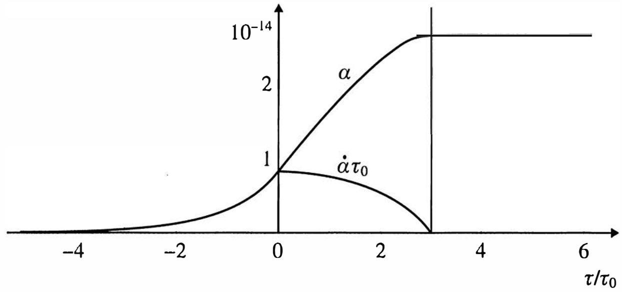


Fig. 9.2 α is the rapidity of an electron and τ_0 is the time taken by a light signal to travel a distance equal to two-thirds of the classical electron radius. A constant force acts from the proper time $\tau_1 = 0$ to the proper time $\tau_2 = 3\tau_0$. The electron, originally at rest, gets a motion (pre-acceleration) before the force acts (after [9.14], courtesy of the American Association of Physics Teachers)

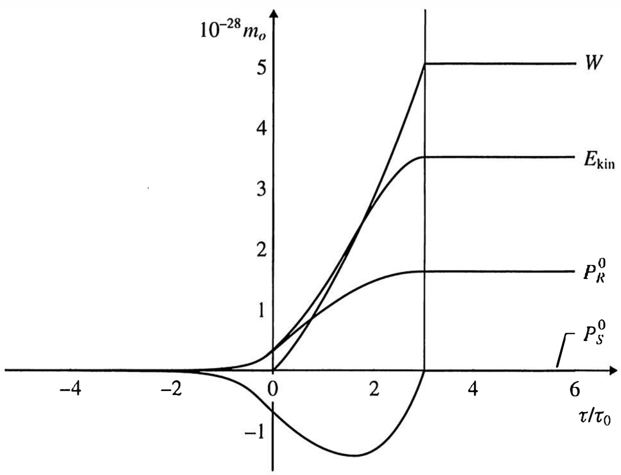


Fig. 9.3 The solution is the same as in Fig. 9.2. The graphs show the kinetic energy E_{kin}, the radiated energy P^0_{R}, the Schott energy P^0_{S}, and the external work W as a function of τ/τ_0. Note that $W = E_{\text{kin}} + P^0_{\text{R}} + P^0_{\text{S}}$. In the pre-acceleration period $P^0_{\text{R}} = E_{\text{kin}}$ (after [9.14], courtesy of the American Association of Physics Teachers)

when we put $\dot{\alpha} = \alpha/\tau_0$. The integral in (9.43) is then solved by introducing $\mathrm{d}\tau = \tau_0\,\mathrm{d}\alpha/\alpha$. We put $\tau_a = -\infty$ and $\tau < \tau_1$. Due to the initial condition $\alpha(-\infty) = 0$ we obtain

$$\Delta\left(m_0 U^\nu\right) = \left(E_{\text{kin}}(\tau), P(\tau)\right) = m_0\left(\cosh\alpha - 1, \sinh\alpha\right), \tag{9.49}$$

$$\Delta P^\nu_{\text{S}} = P^\nu_{\text{S}}(\tau) = m_0\left(-\alpha\sinh\alpha, -\alpha\cosh\alpha\right), \tag{9.50}$$

$$\Delta P^\nu_{\text{R}} = P^\nu_{\text{R}}(\tau) = m_0\left(\alpha\sinh\alpha - \cosh\alpha + 1, \alpha\cosh\alpha - \sinh\alpha\right), \tag{9.51}$$

where

$$\alpha = \tau_0 \mathrm{e}^{\tau/\tau_0} f(\tau_1). \tag{9.52}$$

This leads to

$$\Delta\left(m_0 U^\nu\right) + \Delta P^\nu_{\text{S}} + \Delta P^\nu_{\text{R}} = 0, \tag{9.53}$$

which says that the total increment of the energy and momentum of the system is zero, as it must be since the external force in the interval is zero.

A simple illustration of the above results is obtained by considering the special case where F is constant. We then put $g = F/m_0$, and the solution (9.45)–(9.47) takes the form

$$\tau < \tau_1, \quad \alpha = g\tau_0 \mathrm{e}^{(\tau-\tau_1)/\tau_0}\left(1 - \mathrm{e}^{-(\tau_2-\tau_1)/\tau_0}\right), \tag{9.54}$$

$$\tau_1 < \tau < \tau_2, \quad \alpha = g\tau_0\left(1 - \mathrm{e}^{(\tau-\tau_2)/\tau_0}\right) + g(\tau - \tau_1), \tag{9.55}$$

$$\tau > \tau_2, \quad \alpha = g(\tau_2 - \tau_1). \tag{9.56}$$

The rapidity α and its rate of change times τ_0 are shown graphically in Fig. 9.2. The corresponding curves for the work performed by the external force, $W = \int_{\tau_1}^{\tau} F^0\,\mathrm{d}\tau$, the kinetic energy of the particle, the radiation energy, and the Schott energy, as given in (9.49)–(9.51), are shown in Fig. 9.3.

In order to obtain some intuition about the quantities involved, we may refer to the figures, where we have put $\tau_2 - \tau_1 = 3\tau_0$, and the external force is due to the critical electrical field in air, $E = 2.4\times10^6\,\mathrm{V\,m^{-1}}$. Then for an electron, $g = 4.2\times10^{17}\,\mathrm{m\,s^{-2}}$ and $g\tau_0 = 2.6\times10^{-6}\,\mathrm{m\,s^{-1}}$. In ordinary units, where c is not taken to be 1, the factor $g\tau_0$ in (9.54), say, should be replaced by $g\tau_0/c = 0.88\times10^{-14}$. Hence, according to (9.54) the rapidity in the pre-acceleration period is of the order 10^{-14}, $\alpha(\tau_1) = 0.84\times10^{-14}$, $v(\tau_1) = c\tanh\alpha(\tau_1) = 2.5\times10^{-6}\,\mathrm{m\,s^{-1}}$. To lowest order in α (the next order is of the magnitude 10^{-42}) the expressions (9.49) and (9.50) for the changes of the kinetic energy and the Schott energy, and the emitted radiation energy in the pre-acceleration period reduce to

$$E_{\text{kin}} = m_0(\cosh\alpha - 1) \approx \frac{1}{2} m_0 \alpha^2, \tag{9.57}$$

$$P^0_{\text{S}} = -m_0\alpha\sinh\alpha \approx -m_0\alpha^2, \tag{9.58}$$

$$P^0_R = m_0(1+\alpha\sinh\alpha - \cosh\alpha) \approx \frac{1}{2}m_0\alpha^2\,, \quad (9.59)$$

where α is given by (9.54). These expressions show that radiated energy is approximately equal to the increase of kinetic energy.

9.4 Energy Conservation During Runaway Motion

Runaway acceleration seems to be in conflict with the conservation laws of energy and momentum. The momentum and the kinetic energy of the particle increase even when no force acts upon it. The charged particle even puts out energy in the form of radiation. Where do the energy and the momentum come from?

We shall here show that the source of energy and momentum in runaway motion is the so-called Schott energy and momentum [9.30]. During motion of a charge in which the velocity increases, the Schott energy has an increasingly negative value and there is an increasing Schott momentum directed oppositely to the direction of the motion of the charge.

We shall consider a charged particle performing runaway motion along the x-axis. Introducing the rapidity α of the particle its velocity and acceleration is expressed as in (9.29).

For $F=0$, i.e., for a free particle, the solutions of the LAD equation (9.40) are

1.
$$\dot\alpha = 0, \quad \text{i.e.} \quad \alpha = \text{const.}\,, \quad v = \text{const.}, \quad (9.60)$$
which is consistent with Newton's first law;

2.
$$\dot\alpha = k e^{\tau/\tau_0}\,, \quad k \neq 0\,, \quad \text{i.e.} \quad a \neq 0 \quad (9.61)$$
is the runaway solution.

As pointed out by *Dirac* [9.4] a particle in state 1 or 2 will remain in that state as long as no external force is acting. We shall here consider a particle which is at rest, i.e., in state 1., until it is acted upon by a force $F(\tau)$ pointing in the positive$x$-direction, i.e., we consider a solution to the LAD equation without pre-acceleration. The force acts from τ_1 to τ_2. For $\tau > \tau_2$ the particle is again free.

According to (9.43) $\dot\alpha$ is in the present case given by

$$\dot\alpha(\tau) = -\frac{e^{\tau/\tau_0}}{m_0\tau_0}\int_{-\infty}^{\tau} F(\tau')e^{-\tau'/\tau_0}\,d\tau'\,. \quad (9.62)$$

The integral vanishes for $\tau < \tau_1$, which gives $\dot\alpha = 0$ (and $\alpha = 0$). For $\tau > \tau_2$ the integral is independent of τ and we obtain the runaway motion (9.61). If the integral limit $-\infty$ in (9.62) is replaced by ∞, pre-acceleration is introduced, and the runaway motion disappears.

In the following, we examine (9.62) when the force F has constant value F_0 between τ_1 and τ_2, and is equal to zero outside this interval. The solution of the equation of motion is then

$$\tau < \tau_1\,, \quad \dot\alpha = 0\,, \quad \alpha = 0\,, \quad (9.63)$$

$$\tau_1 < \tau < \tau_2\,, \quad \dot\alpha = \frac{F_0}{m_0} - \frac{F_0}{m_0}e^{\frac{\tau-\tau_1}{\tau_0}}\,, \qquad \alpha = \frac{F_0}{m_0}(\tau-\tau_1) - \frac{F_0\tau_0}{m_0}\left(e^{\frac{\tau-\tau_1}{\tau_0}} - 1\right), \quad (9.64)$$

$$\tau_2 < \tau\,, \quad \dot\alpha = -\frac{F_0}{m_0}\left(e^{-\frac{\tau_1}{\tau_0}} - e^{-\frac{\tau_2}{\tau_0}}\right)e^{\frac{\tau}{\tau_0}}\,, \qquad \alpha = \frac{F_0}{m_0}(\tau_2-\tau_1) - \frac{F_0\tau_0}{m_0}\left(e^{-\frac{\tau_1}{\tau_0}} - e^{-\frac{\tau_2}{\tau_0}}\right)e^{\frac{\tau}{\tau_0}}. \quad (9.65)$$

Equation (9.30) shows a strange aspect of the motion. The quantity $\dot\alpha$ contains two terms. The first expresses the relativistic version of Newton's second law, i.e., $F_0 = d(\gamma m_0 v)/dt$. However, the second term represents a runaway motion *oppositely directed* relative to the external force F_0, a highly unexpected mathematical result. According to (9.64) $\dot\alpha$ and α are oppositely directed relative to F_0 during the entire time interval $\tau_1 < \tau < \tau_2$.

At the point of time $\tau = \tau_2$,

$$\dot\alpha(\tau_2) = \frac{F_0}{m_0}\left(1 - e^{\frac{\tau_2-\tau_1}{\tau_0}}\right), \quad (9.66)$$

$$\alpha(\tau_2) = \frac{F_0}{m_0}\left(\tau_2 - \tau_1 + \tau_0 - \tau_0 e^{\frac{\tau_2-\tau_1}{\tau_0}}\right). \quad (9.67)$$

In order to simplify the expressions we let $\tau_2 - \tau_1 \to 0$ and $F_0 \to \infty$ keeping the product $(\tau_2-\tau_1)\cdot F_0 \equiv P$

constant. We then find the limits

$$\dot{\alpha}(\tau_2) = -\frac{P}{m_0\tau_0}\,, \quad \text{i.e.} \quad a = -\frac{P}{m_0\tau_0}\,, \tag{9.68}$$

$$\alpha(\tau_2) = 0\,, \quad \text{i.e.} \quad v = 0\,. \tag{9.69}$$

In this limit the external force is expressed by a δ function

$$F(\tau) = \delta(\tau - \tau_1)P\,. \tag{9.70}$$

Putting $\tau_1 = 0$ we have the situation: for $\tau < 0$ the particle stays at rest. At $\tau = 0$ it is acted upon by the force

$$F = \delta(\tau)P\,, \tag{9.71}$$

giving the particle an acceleration oppositely directed relatively to the force and a vanishing initial velocity,

$$a(0) = a_0 = -\frac{P}{m_0\tau_0}\,, \quad v(0) = 0\,. \tag{9.72}$$

According to (9.63), (9.64) and (9.72) the motion is as follows,

$$\tau < 0\,, \quad \dot{\alpha} = 0\,, \quad \alpha = 0\,, \tag{9.73}$$

$$\tau > 0\,, \quad \dot{\alpha} = a_0 e^{\frac{\tau}{\tau_0}}\,, \quad \alpha = \tau_0 a_0 e^{\frac{\tau}{\tau_0}} - \tau_0 a_0\,. \tag{9.74}$$

The runaway motion for $\tau > 0$ is accelerated, and the velocity $v = \tanh\alpha$ approaches the velocity of light as an unobtainable limit (Fig. 9.4).

The problem is to explain how this is possible for a particle not acted upon by any external force. It must be possible to demonstrate that the energy and momentum of the particle and its electromagnetic field is conserved, and find the force causing the acceleration. Of essential importance in this connection is the Schott energy and the Schott momentum.

Noting that $\dot{\alpha}$ is the acceleration in the instantaneous inertial rest frame of the particle, we find the energies expressed by the rapidity utilizing, from (9.74), that $\dot{\alpha} = a_0 + \alpha/\tau_0$. The kinetic energy of the particle is

$$E_{\text{kin}} = m_0(\gamma - 1) = m_0(\cosh\alpha - 1)\,. \tag{9.75}$$

The radiation energy is

$$E_{\text{R}} = m_0\tau_0 \int_0^\tau \dot{\alpha}^2 \cosh\alpha\, d\tau = m_0(\alpha\sinh\alpha + a_0\tau_0\sinh\alpha - \cosh\alpha + 1)\,. \tag{9.76}$$

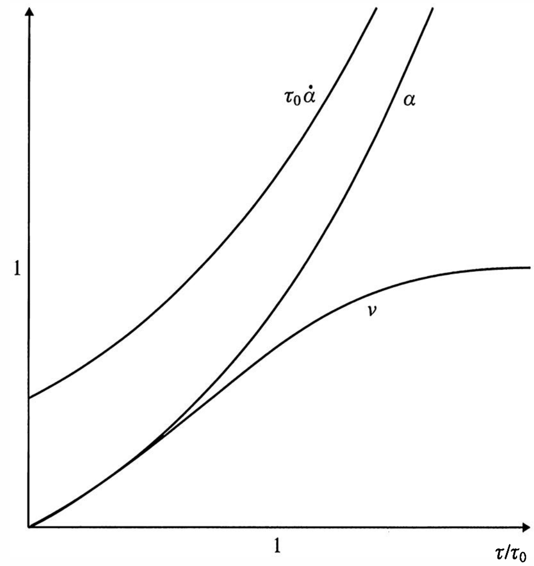


Fig. 9.4 The proper acceleration $\dot{\alpha}$, the velocity parameter α, and the velocity $v = \tanh\alpha$, as functions of the proper time for a particle performing runaway motion, starting from rest with positive acceleration. The quantity τ_0 is the time taken by a light signal to travel a distance equal to two-thirds of the particle's classical radius (after [9.14], courtesy of the American Association of Physics Teachers)

The Schott energy is

$$E_{\text{S}} = -m_0\tau_0\gamma^4 va = -m_0\tau_0\dot{\alpha}\sinh\alpha = -m_0(\alpha + a_0\tau_0)\sinh\alpha\,. \tag{9.77}$$

The sum of the energies is constant and equal to the initial value zero (Fig. 9.5).

Referring to Fig. 9.1 we see that during the runaway motion the sum of the increase of kinetic energy and radiation energy comes from tapping a reservoir of Schott energy which is initially localized very close to the charge. This field energy becomes more and more negative, and the radius of the spherical surface limiting the region with Schott energy increases rapidly. Using (9.26), (9.29), and (9.74) we find that it varies with the proper time of the particle as

$$T - T_{Q2} = \varepsilon e^{\alpha} = \varepsilon \exp\left[|a_0|\exp\left(\frac{\tau}{\tau_0}\right)\right], \tag{9.78}$$

where a_0 is given in (9.72).

Next we consider the momenta. The momentum of the particle is

$$P_{\text{kin}} = m_0\gamma v = m_0\sinh\alpha\,. \tag{9.79}$$

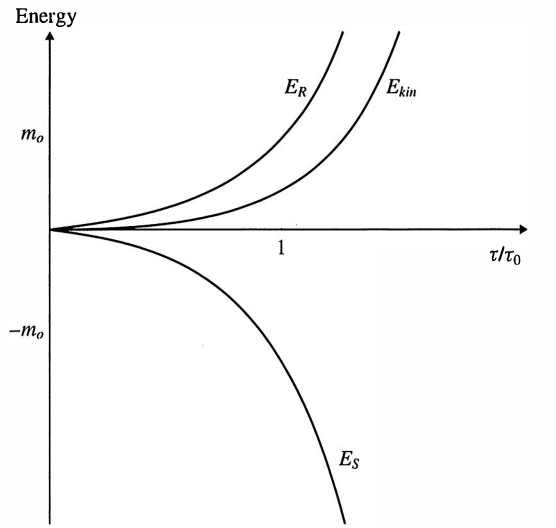


Fig. 9.5 The energies of a particle and its electromagnetic field while the particle performs runaway motion, as functions of τ/τ_0. Here E_{kin} is kinetic energy, E_{R} is radiated energy, and E_{S} is the Schott (or acceleration) energy (after [9.14], courtesy of the American Association of Physics Teachers)

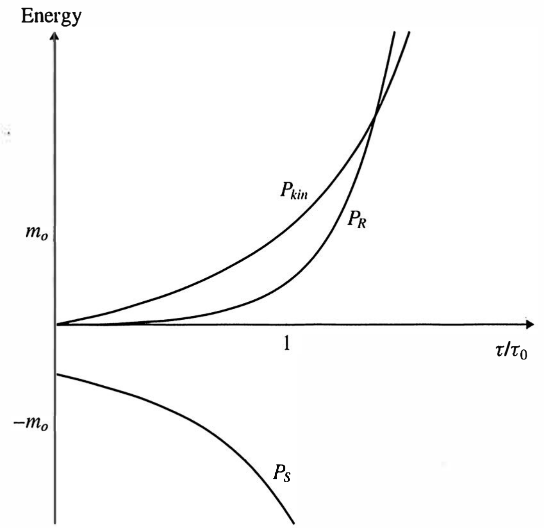


Fig. 9.6 The momentum of a particle and its electromagnetic field while the particle performs runaway motion, as functions of τ/τ_0. Here P_{kin} is kinetic momentum, P_{R} is radiated momentum, and P_{S} is Schott (or acceleration) momentum (after [9.14], courtesy of the American Association of Physics Teachers)

The momentum of the radiation is

$$P_{\text{R}} = m_0 \tau_0 \int_0^{\tau} \dot{\alpha}^2 \sinh \alpha \, d\tau$$
$$= m_0(\alpha \cosh \alpha + a_0 \tau_0 \cosh \alpha - \sinh \alpha - a_0 \tau_0) \,. \quad (9.80)$$

The Schott momentum (acceleration momentum) is

$$P_{\text{S}} = -m_0 \tau_0 \gamma^4 a = -m_0 \tau_0 \dot{\alpha} \cosh \alpha$$
$$= -m_0(\alpha + a_0 \tau_0) \cosh \alpha \,. \quad (9.81)$$

The sum of the momenta is constant and is equal to $-m_0 a_0 \tau_0$, which is the initial Schott momentum (Fig. 9.6).

The forces which are responsible for the increase in the momentum of the particle (internal forces) are the following (for rectilinear motion in general). The radiation reaction force,

$$\Gamma_{\text{R}} = -\frac{dP_{\text{R}}}{dt} = -m_0 \tau_0 \dot{\alpha}^2 \tanh \alpha \,, \quad (9.82)$$

and the acceleration reaction force,

$$\Gamma_{\text{A}} = -\frac{dP_{\text{S}}}{dt} = -\frac{1}{\cosh \alpha} \dot{P}_{\text{S}}$$
$$= m_0 \tau_0 (\ddot{\alpha} + \dot{\alpha}^2 \tanh \alpha) \,. \quad (9.83)$$

The total field reaction force (also called the self-force) is

$$\Gamma = \Gamma_{\text{R}} + \Gamma_{\text{A}} = m_0 \tau_0 \ddot{\alpha} \,. \quad (9.84)$$

By means of (9.82)–(9.84) the forces are shown as functions of τ/τ_0 in Fig. 9.7.

Equation (9.82) shows that the radiation reaction force Γ_{R} is a force that retards the motion, acting like friction in a fluid. The *push* in the direction of the motion is provided by the acceleration reaction force, which is opposite to the change of Schott momentum per unit time. This force is opposite to the direction of the external force, i. e., it has the same direction as the runaway motion.

There is a rather strange point here. We earlier identified the Schott energy as a field energy localized close to the charge [9.7]. Yet, in the present case the Schott

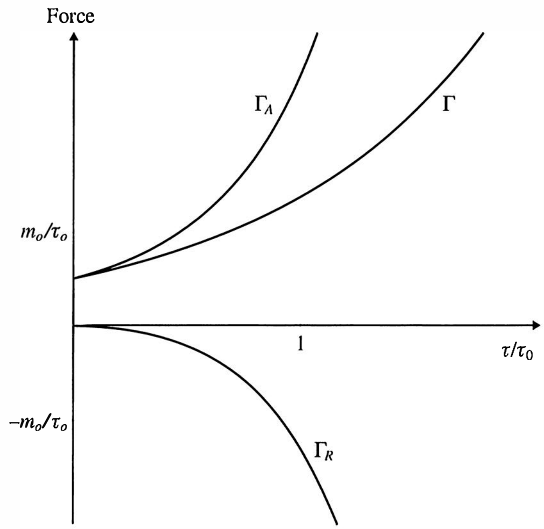


Fig. 9.7 The forces due to the electromagnetic field of a particle acting on the particle while it performs runaway motion, as functions of τ/τ_0. Here Γ_R is the radiation reaction force, Γ_A is the Schott (or acceleration) reaction force. Their sum is the field reaction force, $\Gamma = \Gamma_R + \Gamma_A$ (after [9.14], courtesy of the American Association of Physics Teachers)

momentum is oppositely directed to the motion of the charge. This is due to the fact that the Schott energy is negative. Hence even if the Schott momentum has a direction opposite to that of the velocity of the charge, it represents a motion of negative energy in the same direction as that of the charge.

In general the Schott energy is

$$E_S = -m_0 \tau_0 A^0 , \tag{9.85}$$

and the Schott momentum is

$$\boldsymbol{P}_S = -m_0 \tau_0 \boldsymbol{A} , \tag{9.86}$$

where $(A^0, \boldsymbol{A})$ is the four-acceleration of the particle. From the relation $A^0 = \boldsymbol{v} \cdot \boldsymbol{A}$ we obtain $E_S = \boldsymbol{v} \cdot \boldsymbol{P}_S$. It follows that for rectilinear motion $\boldsymbol{v}$ and $\boldsymbol{P}_S$ are oppositely directed when E_S is negative.

The Schott energy saves energy conservation for runaway motion of a radiating charge. Nevertheless, physicists doubt that runaway motion really exists. There is, however, no doubt that it is a solution of the LAD equation of motion of a charged particle. In this sense it is allowed, but maybe not everything that is allowed is obligatory. The physics equations seem to contain many possibilities that are not realized in our universe. Moreover, we seem to lack a criterion to eliminate those possibilities that do not exist physically. Hence one can only wonder why no runaways have ever been observed or why they could not be used as compact particle accelerators.

9.5 Schott Energy and Radiated Energy of a Freely Falling Charge

The Rindler coordinates (t, x, y, z) of a uniformly accelerated reference frame are given by the following transformation from the coordinates (T, X, Y, Z) of an inertial frame,

$$gt = \operatorname{artanh}\left(\frac{T}{X}\right) , \quad x = \sqrt{X^2 - T^2} , \tag{9.87}$$

with inverse transformation

$$T = x \sinh(gt) , \quad X = x \cosh(gt) . \tag{9.88}$$

Here g is a constant which shall be interpreted physically below.

Using Rindler coordinates, the line element takes the form [9.31]

$$ds^2 = -g^2 x^2 dt^2 + dx^2 + dy^2 + dz^2 . \tag{9.89}$$

In the Rindler frame the nonvanishing Christoffel symbols are

$$\Gamma^x_{tt} = g^2 x , \quad \Gamma^t_{tx} = \Gamma^t_{xt} = \frac{1}{x} . \tag{9.90}$$

The Rindler coordinates are mathematically convenient, but not quite easy to interpret physically. An observer at rest in a uniformly accelerated reference frame in flat spacetime experiences a field of gravity. From the geodesic equation,

$$\frac{du^\mu}{d\tau} + \Gamma^\mu_{\alpha\beta} u^\alpha u^\beta = 0 , \tag{9.91}$$

which is also the equation of motion of a freely moving particle, follows that acceleration of a free particle

instantaneously at rest is

$$\frac{d^2x}{dt^2} = -\Gamma^x_{tt} = -g^2 x\,. \qquad (9.92)$$

Consider a fixed reference point $x =$ constant in the Rindler frame. It has velocity and acceleration

$$V = \frac{dX}{dT} = \tanh(gt)\,, \qquad A = \frac{dV}{dT} = \frac{1}{x}\frac{1}{\cosh^3(gt)}\,, \qquad (9.93)$$

respectively, in the inertial frame. Hence, the acceleration of a reference point $x =$ constant at the point of time $t = 0$ is

$$A(0) = \frac{1}{x}\,. \qquad (9.94)$$

This shows that the coordinate x of the Rindler frame has dimension 1 divided by acceleration and that it is equal to the inverse of the acceleration of the reference point that it represents at the point of time $t = 0$. The physical interpretation of the constant g then follows from (9.92). It represents the acceleration of gravity experienced in the Rindler frame at the reference point having acceleration equal to g relative to the inertial frame at the point of time $t = 0$.

The four-velocity and the four-acceleration of a particle moving along the x-axis are

$$u^\mu = \frac{dx^\mu}{d\tau} = \gamma(1, v, 0, 0)\,, \qquad \gamma = (g^2x^2 - v^2)^{-\frac{1}{2}}\,, \qquad (9.95)$$

$$a^\mu = \frac{du^\mu}{d\tau} + \Gamma^\mu_{\alpha\beta}u^\alpha u^\beta = \gamma^4\left(a + g^2x - \frac{2v^2}{x}\right)(v, g^2x, 0, 0)\,, \qquad (9.96)$$

where $v = dx/dt$ and $a = dv/dt$.

As seen from the expression (9.7) for the Abraham four-force the field reaction force vanishes for a freely moving charge. Hence, such a charge falls with the same acceleration as a neutral particle. It has vanishing four-acceleration and follows a geodesic curve. This is valid in a uniformly accelerated reference frame in flat spacetime, but not in curved spacetime [9.32].

From the expression (9.96) it is seen that in the present case the equation of motion may be written as

$$x\frac{d^2x}{dt^2} - 2\left(\frac{dx}{dt}\right)^2 + g^2x^2 = 0\,. \qquad (9.97)$$

The solution of this equation for a particle falling from $x = x_0$ at $t = 0$ is

$$x = \frac{x_0}{\cosh(gt)}\,. \qquad (9.98)$$

Hence

$$v = -gx_0\frac{\sinh(gt)}{\cosh^2(gt)}\,, \qquad \gamma = \frac{\cosh^2(gt)}{gx_0}\,. \qquad (9.99)$$

In order to give a correct description of the radiation emitted by a charge valid in an accelerated frame of reference, one has to generalize the usual form of the Larmor formula for the radiated effect $\hat{P}$ valid in the orthonormal basis of an inertial frame,

$$\hat{P} = m_0\tau_0\hat{a}^2\,, \qquad (9.100)$$

where $\hat{a}$ is the acceleration of the charge in an inertial frame. This formula says that an accelerated charge radiates, which is a misleading statement. It sounds as if whether a charge radiates or not, is something invariant that all observers can agree upon. However, that is not the case. An accelerated observer permanently at rest relative to an accelerated charge would not say that it radiates. The covariant generalization of the formula is

$$P_L = m_0\tau_0 A_\mu A^\mu\,. \qquad (9.101)$$

Freely falling charges have vanishing four-acceleration. Hence, this version of Larmor's formula seems to say that charges that are acted upon by nongravitational forces radiate. As mentioned above this is not generally the case. In fact, the formula above is not generally covariant. It is only Lorentz covariant because the components of the four-acceleration are presupposed to be given with reference to an inertial frame in this formula.

Saying that a charge radiates is not a reference-independent statement. This conclusion has been arrived at in different ways [9.33–38]. *M. Kretzschmar* and *W. Fugmann* [9.37, 38] generalized Larmor's formula (9.100) to a form which is valid not only in inertial reference frames, but also with respect to accelerated frames. A consequence of their formula is that a charge will be observed to emit radiation only if it accelerates relative to the observer. Whether it moves

along a geodesic curve is not decisive. A freely falling charge, i. e., a charge at rest in an inertial frame may be observed to radiate, and a charge acted upon by non-geodesic forces may be observed not to radiate.

Hirayama [9.39] recently generalized the Lorentz covariant formula (9.101) to one which is also valid in a uniformly accelerated reference frame. He then introduced a new four-vector which may be called the *Rindler four-acceleration* of the charge. It is a four-vector representing the acceleration of the charge *relative to the Rindler frame*, and has components

$$\alpha^{\mu} = a^{\mu} - \frac{1}{gx^2(gx+v)}(v, g^2x^2, 0, 0)\,, \tag{9.102}$$

where a^{μ} are the components in the Rindler frame of the four acceleration of the charge. For a freely falling charge $a^{\mu} = 0$. The generalized Larmor formula valid in a uniformly accelerated reference frame has the form

$$P = m_0\tau_0 g^2 x^2 \alpha_{\mu}\alpha^{\mu}\,, \tag{9.103}$$

and has been thoroughly discussed by *Eriksen* and *Grøn* [9.31].

We shall now apply this formula to the charge falling freely from $x = x_0$ where it was instantaneously at rest. Then we need to calculate

$$\alpha_{\mu}\alpha^{\mu} = -g^2x^2(\alpha^t)^2 + (\alpha^x)^2\,. \tag{9.104}$$

Inserting the expressions (9.98) and (9.99) for x and v in (9.102), we obtain

$$\begin{aligned} \alpha^t &= \frac{1}{gx_0^2}\mathrm{e}^{gt}\sinh(gt)\cosh^2(gt)\,, \\ \alpha^x &= -\frac{1}{x_0}\mathrm{e}^{gt}\cosh^2(gt)\,, \end{aligned} \tag{9.105}$$

which gives

$$\alpha_{\mu}\alpha^{\mu} = \frac{1}{x_0^2}\mathrm{e}^{2gt}\cosh^2(gt)\,. \tag{9.106}$$

Inserting this into (9.103) we find the power radiated by the freely falling charge

$$P = m_0\tau_0 g^2 \mathrm{e}^{2gt}\,. \tag{9.107}$$

The radiated energy is

$$E_{\mathrm{R}} = \int_0^t P\,\mathrm{d}t = \frac{m_0\tau_0}{2}g(\mathrm{e}^{2gt} - 1)\,. \tag{9.108}$$

One may wonder where this energy comes from. A proton and a neutron will perform identical motions during the fall, although the proton radiates energy and the neutron does not. The answer is: the radiated energy comes from the Schott energy. The Schott energy is given by

$$E_{\mathrm{S}} = -m_0\tau_0 v\alpha^x\,. \tag{9.109}$$

Inserting the expressions (9.99) and (9.105) for v and α^x, respectively, we obtain for the Schott energy as a function of time

$$E_{\mathrm{S}} = -\frac{m_0\tau_0}{2}g(\mathrm{e}^{2gt} - 1)\,. \tag{9.110}$$

This shows that the radiation energy does indeed come from the Schott energy. Again the Schott part of the field energy inside the light front S in Fig. 9.1 becomes more and more negative during the motion, and the region filled with Schott energy which is inside the light front S and outside the particle, initially has a vanishing volume, but increases rapidly in size.

9.6 Noninvariance of Electromagnetic Radiation

F. Rohrlich was one of the first to note the noninvariance of electromagnetic radiation from a point charge against a transformation involving a relative acceleration between two reference frames [9.33]. He considered a uniformly accelerated charge in flat spacetime and concluded that a freely falling observer would see a supported charge in a uniformly accelerated reference frame radiating, and a supported observer would see a freely falling charge radiating. But a supported observer would not see a supported charge radiating, and a freely falling observer would not see a freely falling charge radiating. The same was later noted by *A. Kovetz* and *G. E. Tauber* [9.34], and an explanation for this was given by *D. G. Boulware* [9.36].

The nature of electromagnetic radiation is still a mystery. The wave–particle duality is something which seems to transcend our intuitive understanding. The waves of monochromatic light have infinite extension, but a photon is thought of as something having an exceedingly minute extension with

a smallness only limited by the Heisenberg uncertainty relations.

Also thinking of electromagnetic radiation as a photon gas, and photons as a sort of object which you can detect with your apparatus, it seems exceedingly strange to claim that you can make the object vanish just by changing your state of motion. On the other hand that claim does not sound so impossible if you think of electromagnetic radiation as waves. The waves are a state of oscillation of electric and magnetic fields moving through space with the velocity of light. Maybe they can be transformed away?

That should indeed be possible. Think of a uniformly accelerated charge, radiating out an electromagnetic power. Transforming to the permanent rest frame of the charge the magnetic field vanishes. In this frame the charge does not radiate. Hence, saying that a charge radiates is not a reference-independent statement.

As mentioned in Sect. 9.5 *M. Kretzschmar* and *W. Fugmann* [9.37, 38] and *T. Hirayama* [9.39] deduced a generalized versions of Larmors's formula valid in uniformly accelerated reference frames. The significance of these formulae in connection with energy momentum conservation of a charged particle and its electromagnetic field has been thoroughly analyzed by *Eriksen* and *Grøn* [9.31].

The nonvanishing Christoffel symbols in the Rindler frame are given in (9.90), and the four-velocity and the four-acceleration of a particle moving along the x-axis have components given in (9.95) and (9.96). Transformation by means of (9.87) and (9.88) gives the following components of the four-velocity and four-acceleration in the inertial frame,

$$\begin{aligned} U^{\mu} &= (gxv^t, v^x, 0, 0)\,, \\ A^{\mu} &= (gxa^t, a^x, 0, 0)\,. \end{aligned} \tag{9.111}$$

Inserting $v = a = 0$ in (9.95) and (9.96) we find the four-velocity and four-acceleration of the reference particles in the Rindler frame

$$u^{\mu} = \left(\frac{1}{g}x, 0, 0, 0\right)\,, \quad g^{\mu} = \left(0, \frac{1}{x}, 0, 0\right)\,. \tag{9.112}$$

Using (9.96) we find that the proper acceleration $\hat{a}$ (relative to an instantaneous inertial rest frame) is given by $\hat{a}^2 = a_{\mu}a^{\mu} = A_{\mu}A^{\mu}$, i.e.,

$$\hat{a} = \gamma^3 gx\left(a + g^2x - \frac{2v^2}{x}\right)\,. \tag{9.113}$$

For a particle instantaneously at rest at the point $x = x_1$ we obtain

$$\hat{a} = \frac{a}{g^2x_1^2} + \frac{1}{x_1}\,, \tag{9.114}$$

where $1/x_1$ is the proper acceleration of the point $x = x_1$ in the Rindler frame. The difference $\hat{a} - 1/x_1$ will be denoted by $\tilde{a}$. We obtain

$$\tilde{a} = \frac{a}{g^2}x_1^2\,, \tag{9.115}$$

which may be interpreted as the acceleration of the particle relative to the Rindler system, measured by a standard clock carried by the particle.

According to the analysis of *Kretzschmar* and *Fugmann* [9.37, 38] the generalized Larmor formula as written out in a uniformly accelerated reference frame takes the form

$$P = \frac{2}{3}Q^2\,(gx_1)^2\,\tilde{a}^2\,. \tag{9.116}$$

We shall now consider a freely falling charge in the Rindler frame. It is permanently at rest in the inertial comoving frame. Obviously it does not radiate as observed in this frame. But according to (9.116) it radiates as observed in the Rindler frame. In order to understand this from a field theoretic perspective in a similar way as was obtained with reference to an inertial frame in Sect. 9.2, we may again consider the Teitelboim partition of the field into a generalized Coulomb field I and a radiation field II. Calculating the flow of field energy of these types out of the Rindler section we arrived at in [9.31],

$$P_{\mathrm{I}} = \frac{2}{3}Q^2g\,(v - gx_1)\left[\gamma^2g\,(v - gx_1) + 2a^x\right]\,, \tag{9.117}$$

$$P_{\mathrm{II}} = \frac{2}{3}Q^2\left(\frac{a^x}{\gamma}\right)^2\,. \tag{9.118}$$

The emitted energy per emission time is

$$P = P_{\mathrm{I}} + P_{\mathrm{II}} = \frac{2}{3}\frac{Q^2}{\gamma^2}\left[a^x + \gamma^2g\,(v - gx_1)\right]^2\,, \tag{9.119}$$

where

$$a^x = \gamma^4\left(a + g^2x_1 - 2v^2/x_1\right)g^2x_1^2\,. \tag{9.120}$$

We now apply these formulae to the special case of a freely falling charge in the Rindler frame. Then the four-acceleration vanishes, $a^x = 0$, which gives

$$P_{\mathrm{I}} = \frac{2}{3} Q^2 g^2 \gamma^2 (g x_{\mathrm{I}} - v)^2 , \quad P_{\mathrm{II}} = 0 . \tag{9.121}$$

In this case there is no emission of type II energy, only of type I.

This example shows the inadequacy of the Teitelboim separation with respect noninertial reference frames. In inertial frames radiation is associated with type II energy. However, as is seen from the present results, this is not the case in general. The separation in type I and II energy is based, respectively, on the vanishing and the nonvanishing of the four-acceleration of the charge, which means whether it is in free fall or not. The emission of radiation, on the other hand, depends upon the relative acceleration between the charge and the observer. Only in an inertial frame does the vanishing of the four-acceleration mean that the charge is not accelerated relative to the observer. It should also be noted that since there is a flux of type I energy out of the Rindler sector, type I energy is not a state function of the charge in the Rindler frame, as it is in an inertial frame.

Following *Hirayama* [9.39] we shall now present a separation of the electromagnetic field energy in two types, $\tilde{P}_{\mathrm{I}}$ and $\tilde{P}_{\mathrm{II}}$, making use of a modified acceleration called α. We write $\alpha^x = a^x - \Delta^x$, where Δ^x is a quantity independent of α^x, which is determined from the condition that there shall be no energy of the new type I emitted out of the Rindler system, $\tilde{P}_{\mathrm{I}} = 0$. Inserting $a^x = \alpha^x + \Delta^x$ into (9.121) and selecting the term of second order in α^x gives

$$\tilde{P}_{\mathrm{II}} = \frac{2}{3} Q^2 (\alpha^x/\gamma)^2 . \tag{9.122}$$

Since the total transport of energy out of the sector is independent of the partition used, we have $\tilde{P}_{\mathrm{I}} = P - \tilde{P}_{\mathrm{II}}$. Hence, by means of (9.122) we obtain,

$$\begin{aligned} \tilde{P}_{\mathrm{I}} = \frac{2}{3} \frac{Q^2}{\gamma^2} & \left(2\alpha^x + \Delta^x + \gamma^2 g v - \gamma^2 g^2 x_{\mathrm{I}}\right) \\ & \times \left(\Delta^x + \gamma^2 g v - \gamma^2 g^2 x_{\mathrm{I}}\right) . \end{aligned} \tag{9.123}$$

From the requirement that $\tilde{P}_{\mathrm{I}} = 0$ for all values of α^x follows

$$\Delta^x = \gamma^2 g (g x_{\mathrm{I}} - v) , \tag{9.124}$$

giving

$$\alpha^x = a^x - \gamma^2 g (g x_{\mathrm{I}} - v) , \tag{9.125}$$

and

$$\tilde{P}_{\mathrm{I}} = 0 , \quad \tilde{P}_{\mathrm{II}} = P . \tag{9.126}$$

Here α^x is just the x-component of *the acceleration of the charge relative to the Rindler frame* found by Hirayama using Killing vectors.

The covariant expression of the vector is

$$\begin{aligned} \alpha^\mu = {} & a^\mu - \left(g_\alpha g^\alpha\right)^{\frac{1}{2}} u^\mu - g^\mu \\ & - \left(g_\alpha g^\alpha\right)^{\frac{1}{2}} v_\beta u^\beta v^\mu - v_\alpha g^\alpha v^\mu . \end{aligned} \tag{9.127}$$

Using (9.95), (9.96) and (9.112), we have in our case

$$\begin{aligned} \left(g_\alpha g^\alpha\right)^{\frac{1}{2}} &= \frac{1}{x_{\mathrm{I}}} , \\ v_\beta u^\beta &= -\gamma g x_{\mathrm{I}} , \\ v_\alpha g^\alpha &= \gamma \frac{v}{x_{\mathrm{I}}} , \end{aligned} \tag{9.128}$$

and Hirayama's vector reads

$$\alpha^\mu = a^\mu - \frac{\gamma^2 (g x_{\mathrm{I}} - v)}{g x_{\mathrm{I}}^2} \left(v, g^2 x_{\mathrm{I}}^2, 0, 0\right) , \tag{9.129}$$

or, by means of (9.113),

$$\alpha^\mu = \left[\gamma^4 \left(a - \frac{v^2}{x_{\mathrm{I}}}\right) + \frac{\gamma^2 v}{g x_{\mathrm{I}}^2}\right] \left(v, g^2 x_{\mathrm{I}}^2, 0, 0\right) . \tag{9.130}$$

It follows that $(\alpha^x/\gamma)^2 = g^2 x_{\mathrm{I}}^2 \alpha_\mu \alpha^\mu$, which by means of (9.125) gives

$$P = \tilde{P}_{\mathrm{II}} = \frac{2}{3} Q^2 g^2 x_{\mathrm{I}}^2 \alpha_\mu \alpha^\mu , \tag{9.131}$$

for the field energy produced per coordinate time which leaves the Rindler sector.

It is easily seen that the Hirayama separation is a proper generalization of the Teitelboim separation to accelerated frames, which reduces to the latter in inertial frames. To that end we describe the particle by the coordinate $\bar{x} = x_{\mathrm{I}} - 1/g$. Then the coordinate time for $\bar{x} = 0$ is equal to the proper time. Keeping $\bar{x}$ finite

and letting $g \to 0$ we obtain the limits $x_1 \to \infty$, $gx_1 \to 1$, $ds^2 \to -dt^2 + dx^2$, $\gamma \to (1-v^2)^{-1/2}$. From (9.129) we then find that $\alpha^\mu \to a^\mu$, and from (9.131) that $P \to (2/3)\, Q^2 a_\mu a^\mu$.

Calculating the bound energy in the Rindler frame we find that the total energy of the charge and its field is [9.31]

$$\tilde{U} = \tilde{U}_\mathrm{I} + \tilde{U}_\mathrm{II} = -\frac{1}{2} Q^2 g + g^2 x_1^2 \gamma m_0 + \tilde{E}_\mathrm{S} + \tilde{E}_\mathrm{R} . \tag{9.132}$$

The first term on the right-hand side has no obvious physical interpretation. The second is the mechanical energy of the particle. The third term is the acceleration energy or the Schott energy, when the partition of the field is made according to the acceleration α^μ,

$$\tilde{E}_\mathrm{S} = -\frac{2}{3} Q^2 v \alpha^x . \tag{9.133}$$

The quantity $\tilde{E}_\mathrm{S}$ is analogous to the Schott energy

$$E_\mathrm{S} = -\frac{2}{3} Q^2 V A^X , \tag{9.134}$$

in an inertial system according to the Teitelboim partition. The fourth term is the radiation energy in the α^μ-partition,

$$\tilde{E}_\mathrm{R} = \frac{2}{3} Q^2 \int_{-\infty}^{t_1} g^2 x^2 \alpha_\mu \alpha^\mu \, dt . \tag{9.135}$$

By differentiation upon the proper time of the particle, i. e., $d/d\tau = \gamma(d/dt_1)$, we find the formula

$$\frac{d}{d\tau}\left(\gamma g^2 x_1^2\right) = a^t g^2 x_1^2 = -a_t , \tag{9.136}$$

by which the energy equation (9.132) becomes

$$\frac{d\tilde{U}}{d\tau} = -m_0 a_t + \tilde{\Gamma}_0 , \tag{9.137}$$

where

$$\tilde{\Gamma}_0 = \frac{d}{d\tau}\left(\tilde{E}_\mathrm{S} + \tilde{E}_\mathrm{R}\right) = \frac{2}{3} Q^2 \left(\frac{d\alpha_t}{d\tau} - v_t \alpha_\nu \alpha^\nu \right) . \tag{9.138}$$

The quantity $\tilde{\Gamma}_0$ is interpreted as a component of the Abraham vector in the Rindler frame. Let us compare it with the time component of the corresponding vector Γ^μ given by *Hirayama* [9.39]. From his (9.50) we obtain for the motion in the x-direction,

$$\Gamma_0 = \frac{2}{3} Q^2 \left\{ v^\nu \nabla_\nu \alpha_t - \frac{\alpha_t}{\gamma v x} - v_t \alpha_\nu \alpha^\nu \right\} . \tag{9.139}$$

Inserting

$$v^\nu \nabla_\nu \alpha_t = \frac{d\alpha_t}{d\tau} - \Gamma^\sigma_{t\beta} v^\beta \alpha_\sigma = \frac{d\alpha_t}{d\tau} + \frac{\alpha_t}{\gamma v x} , \tag{9.140}$$

we obtain

$$\Gamma_0 = \frac{2}{3} Q^2 \left\{ \frac{d\alpha_t}{d\tau} - v_t \alpha_\nu \alpha^\nu \right\} , \tag{9.141}$$

which is equal to $\tilde{\Gamma}_0$ as given by (9.141). Thus for the Abraham vector in the Rindler frame we have

$$\Gamma_0 = \tilde{\Gamma}_0 = \frac{d}{d\tau}\left(\tilde{E}_\mathrm{S} + \tilde{E}_\mathrm{R}\right) . \tag{9.142}$$

Note that when $\Gamma_0 = 0$, the radiation energy is supplied by the Schott energy. This is quite similar to the corresponding case in an inertial frame. From [9.6, eq. (3.2)] we then have $\Gamma_T = d/d\tau(E_\mathrm{S} + E_\mathrm{R})$, where $E_\mathrm{S} = (2/3)\, Q^2 A_T$ and $E_\mathrm{R} = \frac{2}{3} Q^2 \int_{-\infty}^{T} A_\mu A^\mu \, dT$.

From the Lorentz invariance of Maxwell's equations it follows that the existence of electromagnetic radiation is Lorentz invariant. The quantum mechanical photon picture of radiation suggests that its existence is generally invariant. However, as we have shown in this section, the equations of classical electrodynamics imply that this is not the case. The existence of radiation from a charged particle is not invariant against a transformation involving reference frames that accelerate or rotate relative to each other. Even if a charge accelerates as observed in an inertial frame, it does not radiate as observed from its permanent rest frame.

D. R. Rowland [9.16] recently explained this in the case of a uniformly accelerated charge in the following way:

> *The electric field lines of the charge in the Rindler frame in which it is at rest lie along the geodesics for photons for that frame of reference. This means that relative to the Rindler frame, the photons emitted by the charge are purely longitudinal, not transverse, meaning that they are virtual rather than real (i. e. radiation) photons.*

9.7 Other Equations of Motion

The problems with the LAD equation of motion of a charged point particle, i. e., pre-acceleration and runaway solutions, have motivated several researchers to propose alternative equations of motion. *R. T. Hammond* [9.40] recently reviewed some proposals for constructing a new equation of motion of a radiating electron. The equation of motion is written in the form (9.6). In the case of the LAD equation the vector Γ^μ is given by (9.7). We now consider a charged particle in an external electromagnetic field as described with reference to an inertial frame. Then the external force is given by $F^\mu = qF^{\mu\sigma}U_\sigma$, where $F^{\mu\sigma}$ are the components of the electromagnetic field tensor.

Another equation that has been much used, is the so-called Landau–Lifschitz (LL) equation of motion [9.41], which is

$$\Gamma^\mu = \tau_0 \left(q\dot{F}^{\mu\sigma}U_\sigma + q^2\left(F^{\mu\gamma}F^\alpha_\gamma U_\alpha + F^{\nu\gamma}F^\alpha_\nu U_\alpha U^\mu\right)\right) . \tag{9.143}$$

In the nonrelativistic limit the LL equation takes the form

$$m_0\dot{\boldsymbol{v}} = \boldsymbol{f}_{\text{ext}} + \tau_0\dot{\boldsymbol{f}}_{\text{ext}} . \tag{9.144}$$

In the absence of an external force this reduces to Newton's first law, and there is now a runaway solution. Moreover, there is no pre-acceleration. Hammond pointed out, however, that in the deduction of this equation one utilizes a condition which in the case of a charge oscillating with frequency Ω takes the form (inserting the velocity of light in this formula),

$$\Omega << (f_{\text{ext}}/m_0c\tau_0)^{\frac{1}{2}} . \tag{9.145}$$

Hammond notes that for weak enough external forces this condition may never be satisfied. He comments further on this [9.40]:

> *Of course in this case the net force is extremely small, but for long times, such as charged particles in a galactic orbit, we see that we cannot even use the LL equation. Thus there is an entire range in which the LL equation seems to fail.*

G. W. Ford and *R. F. O'Connell* constructed a more general equation of motion for a radiating charged particle taking a possible electron structure into account [9.42–46] This equation is

$$\Gamma^\mu = \tau_0 q\left(\left(F^{\mu\sigma}U_\sigma\right)^{\cdot} + U^\mu U_\alpha\left(F^{\alpha\beta}U_\beta\right)^{\cdot}\right) . \tag{9.146}$$

In the nonrelativistic limit this is the same as the nonrelativistic version of the LL equation.

Another approach was developed by *A. D. Yaghjian* [9.47]. He modeled a particle by a shell and assumed that no forces act upon the shell until the time when the force is applied, and obtained

$$\Gamma^\mu = \tau_0\theta(\tau)(\ddot{U}^\mu + U^\mu\dot{U}_\alpha U^\alpha) , \tag{9.147}$$

where $\theta(\tau)$ is the step function. Due to the presence of the step function the equation of motion with this expression for Γ^μ avoids pre-acceleration.

Yet another approach is followed by *Hammond* [9.40, 48]. In [9.40] he considers the nonrelativistic case in one dimension and writes the equation of motion as

$$m_0\dot{v} = f_{\text{ext}} - f , \tag{9.148}$$

where f is the radiation reaction force. Then he assumes that the radiated effect is given by fv. Combining this with Larmor's formula (9.11) he finds

$$fv = m_0\tau_0(\dot{v})^2 . \tag{9.149}$$

Eliminating f from (9.148) and (9.149) he arrives at

$$m_0 v\dot{v} = vf_{\text{ext}} - m_0\tau_0\dot{v}^2 . \tag{9.150}$$

Integration gives

$$\frac{1}{2}m_0v^2 = \int f_{\text{ext}}\,\mathrm{d}x - \int P_{\mathrm{L}}\,\mathrm{d}t . \tag{9.151}$$

Hence, the increase of kinetic energy equals the work done by the external force minus the energy radiated away. Hammond then says that (9.150) is free of the plagues of the LAD equation. However, that is not quite so. There is a reminiscence of the runaway solution. If there is no external force $f_{\text{ext}} = 0$, (9.150) reduces to $v = -\tau_0\dot{v}$ with general solution $v = v_0\mathrm{e}^{-t/\tau_0}$. Thus, there is an exponentially decaying *runaway solution.* It may be noted that a solution of (9.150) of the same form, $v = v_0\mathrm{e}^{-t/2\tau_0}$ is obtained if there is an external friction like force proportional to the velocity, $f_{\text{ext}} = -(m_0/4\tau_0)v$.

9.8 Conclusion

Seemingly there is a problem with energy conservation connected with the LAD equation of motion of a radiating charge in combination with the Larmor formula for the effect of the radiation emitted by an accelerated charged particle, although a general analysis implies energy conservation for a dynamics based upon these equations [9.49]. The equation of motion has runaway solutions in which a charge accelerates and emits radiation even when it is not acted upon by any exterior force. Where does the increase of kinetic energy and radiation energy come from?

In the present article it has been shown how the Schott energy provides both an increase of the kinetic energy of the particle and the energy it radiates. The Schott energy is the part of the electromagnetic field energy which is proportional to the acceleration of the charge, and for nonrelativistic motion of the charge it is localized close to the charge [9.7]. The Schott energy has the curious property that it can become increasingly negative, which makes it possible to use it as a sort of inexhaustible source of energy in the case of runaway motion.

Also the case of a freely falling charge in the gravitational field which exists in a uniformly accelerated reference frame in flat spacetime, is quite strange. The comoving frame of the charge is an inertial frame in which it is permanently at rest. Obviously the charged particle does not radiate in this frame. Nevertheless it radiates as observed in the accelerated frame [9.31]. Again one may wonder: Where does the radiated energy come from? Again the answer is: It comes from the Schott energy.

We have here demonstrated how this comes about by calculating the radiated energy and the Schott energy as functions of time for runaway motion and for freely falling motion in a gravitational field. This provides an interesting application of the LAD equation that may be useful in the teaching of the electrodynamics of radiating charges. It has been shown that it is necessary to take the Schott energy into account in order to avoid apparent energy paradoxes in the theory of radiating charges based on the LAD equation.

The necessity of taking the Schott energy into account for energy-momentum conservation may point to a problem with the LAD equation or the point particle model of a charge. Whereas there is a physical basis for the Schott energy in the electromagnetic field of a point charge, an energy that becomes negative without bound and supplies limitless radiation energy and kinetic energy of runaway solutions may be a sign of the breakdown of the LAD equation [9.40].

References

9.1 H.A. Lorentz: *The Theory of Electrons* (Dover, New York 1952), p. 49, based upon lectures given by Lorentz at Columbia University in 1906

9.2 M. Abraham: *Theorie der Elektrizität*, Vol. 2 (Teubner, Leipzig 1905), Eq. (85)

9.3 M. von Laue: Die Wellenstrahlung einer bewegten Punktladung nach dem Relativitätsprinzip, Ann. Phys. **28**, 436–442 (1909)

9.4 P.A.M. Dirac: Classical theory of radiating electrons, Proc. R. Soc. Lond. **167**, 148–169 (1938)

9.5 W.T. Grandy Jr.: *Relativistic Quantum Mechanics of Leptons and Fields* (Kluwer Academic, Dordrecht 1991) p. 367

9.6 P. Yi: Quenched Hawking radiation and the black hole pair-creation rate, ArXiv: gr-qc/9509031

9.7 E. Eriksen, Ø. Grøn: Electrodynamics of hyperbolically accelerated charges. IV: Energy-momentum conservation of radiating charged particles, Ann. Phys. **297**, 243–294 (2002)

9.8 F. Rohrlich In: Lectures in Theoretical Physics, Vol. 2, ed. by W.E. Brittain, B.W. Downs (Interscience, New York 1960)

9.9 E. Eriksen, Ø. Grøn: On the energy and momentum of an accelerated charged particle and the sources of radiation, Eur. J. Phys. **28**, 401–407 (2007)

9.10 G.A. Schott: On the motion of the Lorentz electron, Philos. Mag. **29**, 49–69 (1915)

9.11 F. Rohrlich: The equations of motion of classical charges, Ann. Phys. **13**, 93–109 (1961)

9.12 C. Teitelboim: Splitting of the Maxwell tensor: Radiation reaction without advanced fields, Phys. Rev. **D1**, 1572–1582 (1970)

9.13 P. Pearle: Classical electron models. In: *Electromagnetism: Paths to Research*, ed. by T. Tepliz (Plenum, New York 1982) pp. 211–295

9.14 Ø. Grøn: The significance of the Schott energy for energy-momentum conservation of a radiating charge obeying the Lorentz–Abraham–Dirac equation, Am. J. Phys. **79**, 115–122 (2011)

9.15 E.G.P. Rowe: Structure of the energy tensor in classical electrodynamics of point particles, Phys. Rev. **D18**, 3639–3654 (1978)

9.16 D.R. Rowland: Physical interpretation of the Schott energy of an accelerating point charge and the ques-

tion of whether a uniformly accelerating charge accelerates, Eur. J. Phys. **31**, 1037–1051 (2010)
9.17 H.L. Pryce: The electromagnetic energy of a point charge, Proc. R. Soc. Lond. **168**, 389–401 (1938)
9.18 C.J. Eliezer, A.W. Mailvaganam: On the classical theory of radiating electrons, Proc. Camb. Philos. Soc. **41**, 184–186 (1945)
9.19 A.O. Barut: *Electrodynamics and Classical Theory of Fields and Particles* (Macmillan, New York 1964) p. 196
9.20 H. Levine, E.J. Moniz, D.H. Sharp: Motion of extended charges in classical electrodynamics, Am. J. Phys. **45**, 75–79 (1977)
9.21 E.J. Moniz, D.H. Sharp: Radiation reaction in nonrelativistic quantum electrodynamics, Phys. Rev. D **15**, 2850–2865 (1977)
9.22 E. Tirapequi: On the Lorenz–Dirac equation for a classical charged particle, Am. J. Phys. **46**, 634–637 (1978)
9.23 N.P. Klepikov: Radiation damping forces and radiation from charged particles, Sov. Phys. Usp. **46**, 506–520 (1985)
9.24 J.L. Anderson: Asymptotic conditions of motion for radiating charged particles, Phys. Rev. D **56**, 4675–4688 (1997)
9.25 E.E. Flanagan, R.M. Wald: Does back reaction enforce the averaged null energy condition in semiclassical gravity?, Phys. Rev. D **54**, 6233–6283 (1996)
9.26 J.A. Heras: Preacceleration without radiation: The nonexistence of preradiation phenomenon, Am. J. Phys. **74**, 1025–1030 (2006)
9.27 E. Eriksen, Ø. Grøn: Does preradiation exist?, Phys. Scr. **76**, 60–63 (2007)
9.28 T.C. Bradbury: Radiation damping in classical electrodynamics, Ann. Phys. **19**(323), 347 (1962)
9.29 G.N. Plass: Classical electrodynamic equations of motion with radiative reaction, Rev. Mod. Phys. **33**, 37–62 (1961)
9.30 E. Eriksen, Ø. Grøn: The significance of the Schott energy in the electrodynamics of charged particles and their fields, Indian J. Phys. **82**, 1113–1137 (2008)
9.31 E. Eriksen, Ø. Grøn: Electrodynamics of hyperbolically accelerated charges. V: The field of a charge in the Rindler space and the Milne space, Ann. Phys. **313**, 147–196 (2004)
9.32 B.S. DeWitt, C.M. DeWitt: Falling charges, Physics **1**, 3–20 (1964)
9.33 F. Rohrlich: The principle of equivalence, Ann. Phys. **22**, 169–191 (1963)
9.34 A. Kovetz, G.E. Tauber: Radiation from an accelerated charge and the principle of equivalence, Am. J. Phys. **37**, 382–384 (1969)
9.35 V.L. Ginzburg: Radiation and radiation friction force in uniformly accelerated motion of a charge, Sov. Phys. Usp. **12**, 565–574 (1970)
9.36 D.G. Boulware: Radiation from a uniformly accelerated charge, Ann. Phys. **124**, 169–188 (1980)
9.37 M. Kretzschmar, W. Fugmann: The electromagnetic field of an accelerated charge in the proper reference frame of a noninertial observer, Nuovo Cim. **B103**, 389–412 (1989)
9.38 W. Fugmann, M. Kretzschmar: Classical electromagnetic radiation in noninertial reference frames, Nuovo Cim. **B106**, 351–373 (1991)
9.39 T. Hirayama: Classical radiation formula in the Rindler frame, Prog. Theor. Phys. **108**, 679–688 (2002)
9.40 R.T. Hammond: Relativistic particle motion and radiation reaction in electrodynamics, Electron. J. Theor. Phys. **7**, 221–258 (2010)
9.41 L.D. Landau, E.M. Lifschitz: *The Classical Theory of Fields* (Pergamon Addison-Wesley, Reading 1971), equation 76.1
9.42 G.W. Ford, R.F. O'Connell: Radiation reaction in electrodynamics and the elimination of runaway solutions, Phys. Lett. A **157**, 217–220 (1991)
9.43 G.W. Ford, R.F. O'Connell: Total power radiated by an accelerated charge, Phys. Lett. A **158**, 31–32 (1991)
9.44 G.W. Ford, R.F. O'Connell: Structure effects on the radiation emitted from an electron, Phys. Rev. A **44**, 6386–6387 (1991)
9.45 R.F. O'Connell: The equation of motion of an electron, Phys. Lett. A **313**, 491–497 (2003)
9.46 J. Heras, R.F. O'Connell: Generalization of the Schott energy in electrodynamic radiation theory, Am. J. Phys. **74**, 150–153 (2006)
9.47 A.D. Yaghjian: *Relativistic Dynamics of a Charged Sphere*, Lecture Notes in Physics (Springer, Berlin, Heidelberg 1992)
9.48 R.T. Hammond: New approach to radiation reaction in classical electrodynamics, arXiv: 0902.4231 (2009)
9.49 A.O. Barut: Lorentz–Dirac equation and energy conservation for radiating electrons, Phys. Lett. **131**, 11–12 (1988)

10. The Nature and Origin of Time-Asymmetric Spacetime Structures

H. Dieter Zeh

Time-asymmetric spacetime structures, in particular those representing black holes and the expansion of the universe, are intimately related to other arrows of time, such as the second law and the retardation of radiation. The nature of the quantum arrow, often attributed to a collapse of the wave function, is essential, in particular, for understanding the much discussed *black hole information loss paradox*. This paradox assumes a new form and can possibly be avoided in a consistent causal treatment that may be able to avoid horizons and singularities. The master arrow that would combine all arrows of time does not have to be identified with a direction of the formal time parameter that serves to formulate the dynamics as a succession of global states (a trajectory in configuration or Hilbert space). It may even change direction with respect to a fundamental *physical* clock such as the cosmic expansion parameter if this was formally extended either into a future contraction era or to negative *pre-big-bang* values.

10.1 The Time Arrow of Gravitating Systems

Since gravity is attractive, most gravitational phenomena are asymmetric in time: objects fall down or contract under the influence of gravity. In General Relativity, this asymmetry leads to drastically asymmetric spacetime structures, such as future horizons and future singularities, which would occur, in particular, in black holes. However, since the relativistic and nonrelativistic laws of gravitation are symmetric under time reversal, all time asymmetries must arise as consequences of specific (only seemingly normal) *initial* conditions, for example a situation of rest that can be prepared by means of other arrows of time, such as friction. Otherwise, conclusions like gravitational contraction would have to apply in both directions of time. Indeed, the symmetry of the gravitational laws does allow objects to be thrown up, where their free motion could in principle *end* by another external intervention, or the existence of so called white holes, which would have to contain past singularities and past horizons.

The absence of past horizons and past singularities from our universe (except for a very specific big bang singularity) must be regarded as a time asymmetry characterizing our global spacetime (Sects. 10.2 and 10.4), while Einstein's field equations would not only admit the opposite situation (for example, local past singularities), but also many solutions with mixed or undefined arrows of time – including closed time-like curves and nonorientable spacetimes. Therefore, the mere possibility of posing an *initial* condition is exceptional in general relativity from a general point of view. I will here not discuss such mathematically conceivable solutions that do not seem to be realized in nature, but instead concentrate on models that come close to our universe – in particular those which are globally of Friedmann type. A specific arrow characterizing a Friedmann universe is given by its expansion (unless this would be reversed at some time of maximum extension – see Sect. 10.4).

In many cases, nongravitational arrows of time remain relevant for the evolution of gravitating bodies even *after* the latter have been prepared in an appropriate initial state. This applies, in particular, to strongly gravitating objects, such as stars, whose evolution is essentially controlled by thermodynamics (emission of

heat radiation into the cold universe). The relation between the electrodynamic and thermodynamic arrows (retardation and the second law, respectively) [10.1, Chap. 2] is quite obvious in this case.

Gravitating systems are nonetheless thermodynamically unusual in possessing negative specific heat [10.1, Chap. 5]. This means, for example, that stars become hotter when losing energy through emitting heat, and that satellites accelerate as a consequence of friction in the earth's atmosphere. It can best be understood by means of the virial theorem, which states in its nonrelativistic limit that, for all forces varying with the second negative power of distance (that is, gravitational and Coulomb forces), bound states have to obey the relation $\overline{E}_{\text{pot}} = -2\overline{E}_{\text{kin}}$, where the overbar means averaging over (quasi) periods of time. Therefore,

$$\begin{aligned} E &= E_{\text{pot}}(t) + E_{\text{kin}}(t) = \overline{E}_{\text{pot}} + \overline{E}_{\text{kin}} = \tfrac{1}{2}\overline{E}_{\text{pot}} \\ &= -\overline{E}_{\text{kin}} \propto -T\,. \end{aligned} \tag{10.1}$$

In order to maintain a stable state, these systems must gain from gravitational contraction twice the energy they are losing by radiation or by friction. Nonrelativistically, this negative heat capacity could be bounded by means of other (repulsive) forces that become relevant at high densities, or by the Pauli principle, which controls the density of electrons in white dwarf stars or solid planets, for example. Relativistically, even these limits will break down at a certain mass, since (1) relativistic degeneracy must ultimately lead to the creation of other particles, while (2) the potential energy of repulsive forces will itself gravitate, and for a sufficiently large mass overcompensate any repulsion. Therefore, it is the thermodynamic arrow underlying thermal radiation and the accretion of matter that requires evolution of gravitating systems toward the formation of black holes. Classically, black holes would thus define the final states in the observable evolution of gravitating systems.

10.2 Black Hole Spacetimes

The metric of a spherically symmetric vacuum solution for nonzero mass is shown in Fig. 10.1 in the Kruskal coordinates u and v. This diagram represents the uniquely completed Schwarzschild metric in the form

$$\begin{aligned} \mathrm{d}s^2 = {} & \frac{32M^2}{r}\mathrm{e}^{-r/2M}(-\mathrm{d}v^2 + \mathrm{d}u^2) \\ & + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\,, \end{aligned} \tag{10.2}$$

where the new coordinates u and v are in the external region ($r > 2M$) related to the conventional Schwarzschild coordinates r and t by

$$u = \mathrm{e}^{r/4M}\sqrt{\frac{r}{2M} - 1}\,\cosh\left(\frac{t}{4M}\right), \tag{10.3a}$$

$$v = \mathrm{e}^{r/4M}\sqrt{\frac{r}{2M} - 1}\,\sinh\left(\frac{t}{4M}\right). \tag{10.3b}$$

Each point in the diagram represents a sphere with surface $4\pi r^2$. Note that r and t interchange their roles as space and time coordinates for $r < 2M$, where $2M$ is the Schwarzschild radius. All parameters are given in Planck units $\hbar = G = c = 1$.

As nature seems to provide specific initial conditions in our universe, it may thereby exclude all past singularities, and hence all past event horizons. This initial condition would immediately eliminate the Schwarzschild–Kruskal vacuum solution that is shown in the figure, but we may instead consider the future evolution of a spherically symmetric mass distribution initially at rest, such as a dust cloud. It would classically collapse freely into a black hole, as quantitatively described by the Oppenheimer–Snyder scenario [10.2] (see the left part of Fig. 10.2). The vacuum solution (10.2) is then valid only outside the surface of the dust cloud, but this surface must, according to a classical description, fall through the arising horizon at some finite proper time, and a bit later onto the future singularity.

For a cloud of interacting gas molecules, this gravitational collapse would be thermodynamically delayed by the arising pressure, as indicated in Sect. 10.1. Gravitational radiation would lead to the loss of any kind of macroscopic structure, while whatever remains would become unobservable to an external observer. Although thermodynamic phenomena control the loss of energy by radiation during most of the time, the asymmetric absence of *past* singularities represents a fundamental cosmological initial condition. However, a conceivable white hole initiated by a past singularity that *completely* represented a time-reversed black hole would even require anti-thermodynamics and coherently incoming advanced radiation. So one may suspect that

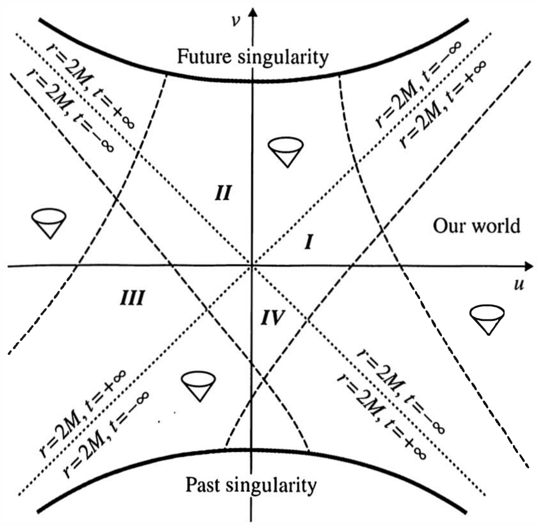


Fig. 10.1 Complete formal continuation of the Schwarzschild solution by means of unique Kruskal coordinates. Quadrants *I* and *II* represent external and internal parts, respectively, of a classical black hole. *III* is another asymptotically flat region, while *IV* would describe the interior of a *white hole.* In this diagram, fixed Schwarzschild coordinates r and t are represented by hyperbola and straight lines through the origin, respectively. Proper times of local objects could start at $t = -\infty$ in *I* or at $t = +\infty$ in *III* (both at $r = \infty$, or at $r = 0$ on the past singularity in *IV*, while they must end at $t = +\infty$ or $-\infty$ in *I* or *III*, respectively, or at a second singularity with the coordinate value $r = 0$ in *II*. On time-like or light-like curves intersecting one of the horizons at the Schwarzschild radius $r = 2M$, the value of the coordinate t jumps from $+\infty$ to $-\infty$ at the rim of quadrant *I*, or from $-\infty$ to $+\infty$ at the rim of quadrant *III*, where t decreases in the physical time direction

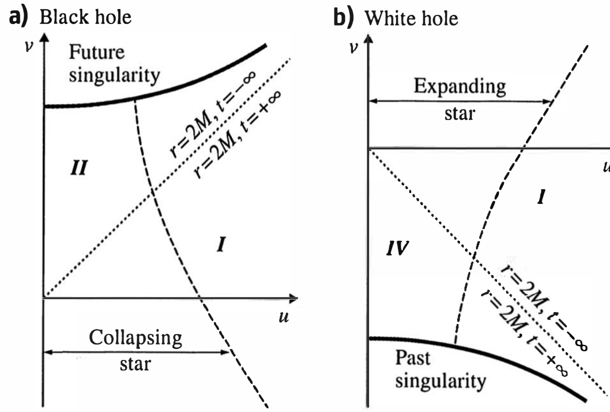


Fig.10.2a,b Oppenheimer–Snyder-type spacetimes of **(a)** a black and **(b)** a white hole

all these various arrows are related to one another, thus defining a common *master arrow of time.*

Since it would require infinite Schwarzschild coordinate time for an object to reach the horizon, any message it may send to the external world would not only be extremely redshifted, but also dramatically delayed. The message could reach a distant observer only at increasingly later stages of the universe. (An apparatus falling into a galactic size black hole could even send messages for a considerable length of *proper* time before it would approach the horizon.) So all objects falling into the black hole must disappear from the view of mortal external observers and their descendents, even though these objects never seem to reach the horizon according to their rapidly weakening, but in principle still arriving signals. The only asymptotically remaining properties of the black hole are conserved ones that have early enough caused effects on the asymptotic metric or other asymptotic fields, namely angular momentum and electric charge. This time-asymmetric consequence is known as the *no-hair theorem* for black holes. During cosmological times, a black hole accumulating ionized interstellar matter may even lose its charge and angular momentum, too, for statistical and dynamical reasons [10.3]. Only its mass and its center of mass motion would then remain observationally meaningful. A black hole is usually characterized by its center of mass system and its long-lasting properties, namely its mass M, charge Q, and angular momentum J, in which case its *Kerr–Newman metric* is explicitly known. The internal topological structures of these metrics for $J \neq 0$ and/or $Q \neq 0$ are radically different from that of the Kruskal geometry in Fig. 10.1, thus raising first doubts in the physical relevance of their formal continuations inside the horizon.

It is important, though, to keep in mind the major causal structure of a black hole: its interior spacetime region *II* would never enter the past of an external observer, that is, it will never become a *fact* for him or her. While the whole exterior region $r > 2M$ can be completely foliated in terms of what will be called *very nice* space-like slices according to increasing Schwarzschild or similar time coordinates with $-\infty < t < +\infty$, the interior can then be regarded as its *global future* continuation beyond the event horizon, where increasing time can be labeled by the Schwarzschild coordinate r decreasing from $r = 2M$

to $r = 0$. This structure must be essential for all causal considerations that include black holes. In this purely classical scenario, the internal state of a black hole would be completely determined by the infalling matter, which could even depend on our free decisions about what to drop into a black hole. Nonetheless, all properties of this infalling matter would irreversibly become *irrelevant* for external observers – a term that is also used to define a generalized concept of coarse graining required for the concept of physical entropy in statistical thermodynamics [10.1, Sect. 3.2].

10.3 Thermodynamics and Fate of Black Holes

In the classical picture described above, a black hole would represent a perfect absorber at zero temperature. This picture had to be corrected when *Bekenstein* and *Hawking* demonstrated [10.4, 5], the latter by explicitly taking into account quantum fields other than gravity, that a black hole must possess finite temperature and entropy proportional to its surface gravity κ and surface area A, respectively,

$$T = \frac{\hbar\kappa}{2\pi k_B} \rightarrow \frac{\hbar c^3}{8\pi G k_B}\frac{1}{M}, \quad (10.4a)$$

$$S = \frac{k_B c^3 A}{4\hbar G} \rightarrow \frac{4\pi k_B G}{\hbar c}M^2 . \quad (10.4b)$$

Here, κ and A are known functions of M, Q, and J, while the explicit expressions given on the right-hand side of the arrow hold for Schwarzschild black holes ($Q = J = 0$) and with respect to spatial infinity (that is, by taking into account the gravitational redshift). This means, in particular, that a black hole must emit thermal radiation (Hawking radiation) proportional to $T^4 A$ according to Stefan–Boltzmann's law, and therefore, that it lives only for a limited time of the order $10^{65}(M/M_{sun})^3$ yr. For astrophysical objects this is many orders of magnitude more than the present age of the universe of about 10^{10} yr, but far less than any Poincaré recurrence times for macroscopic systems.

Even these large evaporation times will begin to *count* only after the black hole has for a very long time to come *grown* in mass by accreting matter [10.6] – at least until the cosmic background temperature has dropped below the very small black hole temperature by means of the growing Hubble redshift. Although evaporation times are thus extremely large, all radiation would causally always precede a genuine horizon. Schwarzschild times represent proper times of distant observers in the rest frame of the black hole, but their corresponding simultaneities may be consistently continued inward while remaining outside the horizon to form complete time coordinates for the whole external region *I*. According to their construction, they would then all have to include the center of the collapsing matter at a prehorizon stage. However, the horizon and interior region *II* could never form if the black hole's energy was indeed radiated away *before* any matter has arrived at the expected horizon in the sense of these simultaneities. So what happens to the infalling matter and, in particular, to nonlocal quantum states in this description?

Schwarzschild simultaneities are counterintuitive. For example, one may use time translation invariance of the external region of the Kruskal-type diagram (Figs. 10.1 or 10.2) in order to define the time coordinate $v = t = 0$ to coincide with an external time close to the peak of the Hawking radiation (in the very distant future from our point of view). Assuming that one can neglect any quantum uncertainty of the metric (which must in principle arise in quantum gravity) for this purpose, all infalling matter that had survived the radiation process so far would at this coordinate time $v = 0$ be in the very close vicinity of the center because of the extreme Lorentz contraction on this simultaneity with respect to the rest system of the infalling matter. Therefore, this simultaneity represents quite different proper times for the various parts of infalling matter even for a collapsing homogeneous dust cloud – and even more so for later infalling things. Most of the black hole's initial energy must already exist in the form of outgoing Hawking radiation at this time, and may even have passed any realistic external observer. If something happens that can become relevant to an external observer (such as the creation of Hawking radiation), it must happen outside the horizon because of relativistic causality.

Black hole radiation is again based on the radiation arrow of retardation, but its conventional formulation also depends on a quantum arrow that is responsible for the statistical interpretation. A pure quantum state gravitationally collapsing toward a black hole would accordingly decay into a collection of various possible decay fragments (mainly photons), described by a sta-

tistical ensemble of all their emission times – similar to a series of unread measurements or to the decay of a highly excited quantum state of a complex object [10.7, 8]. An *apparent* ensemble can be defined even for a resulting *pure* state (according to a unitary description) by means of some physically relevant coarse graining. In quantum theory, one usually neglects in this sense (that is, one regards as irrelevant for the future) the entanglement between all possible decay products and the phase relations between all their decay times. This coarse graining does not only formally justify the concept of growing *physical* entropy in spite of a pure global state [10.1, Sect. 3.2], but also the phenomenon of decoherence. In contrast to the global ensemble entropy that would be conserved under unitary dynamics (and vanishes for a pure state), physical entropy is defined as an extensive quantity, that is, in accord with the concept of an entropy density that neglects information about correlations. The major difference between the decay of highly excited states of normal matter and the evaporation of black holes is that the latter's unitary dynamics is not explicitly known (and occasionally questioned to apply).

The thus described situation is nonetheless much discussed as an *information loss paradox for black holes* [10.9–14]. Its consequences are particularly dramatic if one *presumes* the existence of a black hole interior region that would necessarily arise in the absence of Hawking radiation, since matter (and the *information* it may represent) can then not causally escape any more. This questionable presumption is often introduced by using *nice slices* that are defined to avoid the singularity but may, in contrast to our *very nice slices*, intersect the thus also presumed horizon. A unitary description means, however, that the information defining the initial pure state is partly transformed into nonlocal entanglement (formally analogous to the statistical correlations arising in deterministic Boltzmann collisions). In the quantum case, unitarity leads to a *superposition* of *many worlds* which remain dynamically autonomous, and which may include different versions of the *same* observers (thus physically justifying the concept of decoherence). The replacement of this superposition by an ensemble of many *possible* worlds according to a fundamental statistical interpretation (a collapse of the wave function) would objectively and irreversibly annihilate the information contained in their relative phases, thus introducing a fundamental dynamical time asymmetry. Recall that the Oppenheimer–Snyder model, on which the nice slices are based, precisely neglects the local energy loss due to Hawking radiation. Although the (*back*) reaction of the metric in response to radiation loss may in principle require quantum gravity, my argument about the nonexistence of a horizon is here only based on the local conservation of momentum-energy in a situation where it does not have to be questioned.

Instead of assuming an initial vacuum when calculating the creation of Hawking radiation close to the horizon, one should therefore take into account the presence of infalling matter, in which case some kind of internal conversion might lead to its annihilation. A similar scenario has recently been postulated as a novel kind of physics close to the horizon (called a *firewall*) [10.15]. While this firewall is meant to prevent an observer from remaining intact when falling in, it should according to my earlier proposal objectively convert *all* potentially infalling matter into radiation (see the first version of this paper, available under arXiv:1012.4708v1). Note that the *local* Bekenstein–Hawking temperature diverges close to the horizon, and therefore must lead to the creation of all kinds of particle–antiparticle pairs. As long as such an internal conversion cannot be excluded, there is no reason to speculate about black hole remnants, trapping horizons, superluminal tunneling, or a *fundamental* violation of unitarity that would go beyond decoherence (that is, beyond a mere dislocalization of superpositions) [10.16–18]. The concept of complementarity (in the Copenhagen interpretation) would apply to different *potential* measurements by the same observer, but not to actual measurements to be performed by different ones (including *Wigner's friends*), who would always agree on an objective outcome. Unitarity can only apply to the global *bird's perspective* that includes all Everett branches, while it is incompatible with any kind of *double entanglement* [10.19].

What might remain as a *remnant* according to this semiclassical description of black hole evolution on very nice slices is a *massless pointlike* curvature singularity, since the Riemann tensor of the Schwarzschild metric is proportional to M/r^3, and hence diverges for $r = 2M \to 0$. Evidently, this singularity signals a breakdown of the semiclassical description. Quantum gravity would require a boundary condition for the timeless Wheeler–DeWitt wave function, which cannot distinguish between past and future singularities (Sects. 10.4 and 10.5). This might lead to an effective final condition affecting black holes *from inside* in an anticausal manner [10.1, Sect. 6.2.3]. For example, any inward-directed (hence virtual) negative energy radiation compensating the emission of Hawking radiation could in this way *recohere* the effective black hole state in or-

der to lower its entropy in accordance with both the mass loss and Bekenstein's relation (10.4b). This retrocausality could even affect the nature of the outgoing Hawking radiation. The conventional classical continuation of the metric beyond the horizon according to a *no-drama scenario* may simply be too naive.

Note that the concept of an S-matrix is also unrealistic for macroscopic objects, such as black holes. Because of their never-ending essential interaction with their environments, they can never become asymptotically isolated (the reason for their permanent decoherence). The extreme lifetimes of black holes mean, however, that the information loss problem is rather academic at any rate: any apparently lost information would remain irrelevant for at least the next 10^{65} yr, and it could hardly ever be exploited even if it finally came out in the form of entangled radiation (representing a huge superposition of *many worlds*). The concept of a *Page time* [10.20], at which the entanglement between the residual black hole and its emitted radiation would be maximal, can therefore not have any physical consequences for the remaining black hole.

Several physicists (including myself) used to see a problem in the equivalence principle, which seems to require that observers or detectors freely falling into the black hole do *not* register any black hole radiation. Some even concluded that the mass-loss of black holes, too, must be observer dependent (the already mentioned *black hole complementarity*). However, this conclusion appears to be wrong. While the equivalence between a black hole and a uniformly accelerated detector (as regards their radiation) must indeed apply to the local laws, it can in general *not* do so for their boundary conditions. An observer or detector fixed at some distance from the black hole would not be immersed in *isotropic* heat radiation, since this radiation comes from the black hole surface (or a region close to it), which would cover the whole sky only for an observer very close to the horizon. Even if the infalling detector does not register the radiation at all, the latter's effect on fixed detectors, or its flux through a fixed sphere around the black hole, must exist objectively, just as the clicks of an accelerated detector in an inertial vacuum (attributed to Unruh radiation) can be observed by an inertial observer, too. Therefore, both observers would agree that the energy absorbed by the accelerated detector must be provided by the rocket engine and, analogously, that the Hawking net flux of energy requires an observer-independent mass loss of the black hole. The infalling observer would furthermore have to regard the clicks of fixed detectors as occurring in an extreme quick motion movie with respect to his proper time, and therefore as being caused by an extremely strong outward flux of energy in his reference frame. For the same reason, matter at the rim of a collapsing dust cloud can at large Schwarzschild times not experience any gravitational field as there are no net sources for it inside its present radius any more. So it can never cross a horizon. In this way, the phenomenon of black holes from the point of view of external observers is consistent with the fate of an infalling observer, who may either soon in his proper time have to be affected himself by the internal conversion process, or otherwise have to experience the black hole surface very rapidly shrinking and disappearing before he arrives. (Note, however, that the concept of an event horizon changing in time appears ill-defined in principle, since a horizon is already a spacetime concept.)

If the observer could survive the internal conversion process, he would have traveled far into the cosmic future in a short proper time because of the extreme time dilation close to the would-be horizon. On the other hand, no theory that is compatible with the equivalence principle can describe baryon number nonconservation in the absence of a singularity, although all symmetries can in principle be broken by the effective nonunitarity characterizing the dynamics of an *individual* Everett branch (an *observed quantum world*). This last remark might also be relevant for the above-mentioned possibility of anticausality (recoherence) required by an apparent future condition that would be in accord with a timeless Wheeler–DeWitt equation (Sect. 10.5); recoherence would have to include a reunification of different Everett worlds.

Roger Penrose compared black hole entropy numerically with that of matter in the universe under normal conditions [10.21, 22]. Since the former is proportional to the square of the black hole mass according to (10.4b), macroscopic black hole formation leads to a tremendous increase of entropy. As thermodynamic entropy is proportional to the particle number, it is dominated in the universe by photons from the primordial cosmic radiation (whose number exceeds the baryon number by a factor of 10^9). If our observable part of the universe of about 10^{79} baryons consisted completely of solar mass black holes, it would possess an entropy of order of 10^{98} (in units of k_B^{-1}), that is, 10^{10} times as much as the present matter entropy represented by 10^{88} photons. Combining all black holes into one huge one would even raise this number to 10^{121}, the highest conceivable entropy for this (perhaps partial) universe unless its volume increased

tremendously [10.3, 6, 23]. If entropy is indeed a measure of probability, any approximately homogeneous matter distribution would be extremely improbable except for densities much lower than at present (at a very late stage of an eternally expanding universe). Therefore, the homogeneity of the initial universe is usually regarded as the *fundamental improbable initial condition* that explains the global master arrow of time if statistical reasoning is applicable toward the future (Sect. 10.4). However, its relationship to the thermodynamically important condition of absent or *dynamically irrelevant* nonlocal initial correlations (or entanglement in the quantum case) seems to be not yet fully understood. If the two entropy concepts (black hole and thermodynamic) are to be compatible, the entropy of the final (thermal) radiation must be greater than that of the black hole, while the latter has to exceed that of any kind of infalling matter.

10.4 Expansion of the Universe

The expansion of the universe is a time-asymmetric process, but in contrast to most other arrows, it forms an individual phenomenon rather than a whole class of similar ones, such as black holes, radiation emitters, or steam engines. It may even change its direction at some time of maximum extension, although present astronomical observations may indicate that the expansion will last forever. A homogeneous and isotropic Friedmann universe is described by the dynamics of the expansion parameter $a(t)$ according to the time-symmetric *energy theorem*

$$\frac{1}{2}\left(\frac{1}{a}\frac{\mathrm{d}a}{\mathrm{d}t}\right)^2 = \frac{4\pi}{3}\rho(a) + \frac{\Lambda}{6} - \frac{k}{2a^2}\,, \qquad (10.5)$$

where ρ is the energy density of matter, Λ the cosmological constant, and $k = 0, \pm 1$ the sign of the spatial curvature. The value of the formal *total energy* (the difference of both sides of the equation) is fixed and vanishes in general-relativistic cosmology. Penrose's entropy estimates demonstrate that the homogeneity assumed in (10.5) is extremely improbable from a statistical point of view. Therefore, it must be highly unstable over cosmological times (in spite of being dynamically consistent) under the influence of gravity.

In accordance with a homogeneous initial matter distribution, Penrose postulated that free gravitational fields vanished exactly at the Big Bang. These free fields are described by the *Weyl tensor*, that is, the trace-free part of the curvature tensor. The trace itself (the Ricci tensor) is locally fixed by the stress–energy tensor of matter by means of the Einstein field equations. The Weyl tensor, on the other hand, is analogous to the divergence-free part of the electrodynamic field tensor $F^{\nu\mu}$, whose divergence $\partial_\mu F^{\nu\mu}$ (the trace of the tensor of its derivatives) is similarly fixed by the charge current j^ν. Therefore, the *Weyl tensor hypothesis* is analogous to the requirement of the absence of any free initial electromagnetic radiation, a condition that would leave only the retarded electromagnetic fields of all past sources in the universe. This universal retardation of radiation had indeed been proposed *as a law* by Planck (in a dispute with *Boltzmann*) [10.24], and later by *Ritz* (in a dispute with *Einstein*) [10.25], in order to *derive* the thermodynamic arrow from the law. Here, Boltzmann and Einstein turned out to be right, since the observed retardation is itself a causal consequence of the presence of thermodynamic absorbers [10.1, Chap. 2] – cosmologically including the absorber formed by the radiation era, which would not allow us to observe any conceivable primordial electromagnetic radiation. In contrast, the early universe seems to be transparent to *gravitational* radiation, possibly including that which might have been created with the Big Bang.

Note that the low entropy and corresponding homogeneity of the universe can *not* be explained by an early cosmic inflation era (as has occasionally been claimed) if this inflation is deterministic and would thus conserve ensemble entropy.

Although our universe may expand forever, the idea of its later recontraction is at least conceptually interesting. *Thomas Gold* first argued that the low entropy condition should not be based on an absolute direction of time, and hence be valid at a conceivable Big Crunch as well [10.26]. The latter would then be observed as another Big Bang by observers living during the formal contraction era, while local (black hole) future singularities would be excluded similarly as white holes. Gold's scenario would not only require a transition era without any well-defined arrow in our distant future – it would also pose serious consistency problems, since the extremely small initial probability for the state of the universe would have to be squared if

the two conditions were statistically independent of one another [10.27]. If nonetheless true, it would have important consequences for the fate of matter falling into massive black holes. If such black holes survived the mentioned thermodynamic transition era at the time of maximum extension because of their long evaporation times (Sect. 10.3), they would according to the global dynamics enter an era with reversed arrows of time. However, because of the transparency of the late universe to light, they would *receive* coherent advanced radiation from their formal future even before that happens. This advanced radiation must then *retro-cause* such massive black holes to expand again in order to approach a state of homogeneity in accordance with the final condition [10.28]. In mathematical terms, their horizon is not *absolute* in this case even in the absence of any black hole evaporation.

A reversal of the arrow of time may not only occur in the distant future, but it may also have occurred in the past. Several *pre-big-bang* scenarios have been discussed in novel and as yet speculative theories. Usually, one thereby identifies the direction of the formal time parameter with the direction of the physical arrow of time. For example, according to arguments first used in loop quantum gravity [10.29], the configuration space for Friedmann-type universes may be doubled by interpreting formally negative values of the cosmic expansion parameter a as representing negative volume measures. The cosmic dynamics can then be continued backward in time beyond the Big Bang into its mirror image by *turning space inside out* (turning right-handed triads into left-handed ones) while going through $a = 0$ even in a classical picture. For this purpose, the classical dynamical description (10.5) would have to be modified close to the otherwise arising singularity at $a = 0$ – as it is indeed suggested by loop quantum gravity. However, if the boundary conditions responsible for the arrow of time are still assumed to apply at the situation of vanishing spatial volume, the arrow would formally change direction, and $|a|$ rather than a would represent a physical cosmic clock. Observers on both temporal sides of the Big Bang could only remember events in the direction toward $a = 0$. Another possibility of avoiding the singularity is a repulsive force acting at small values of a [10.30, 31], which would lead to a Big Bounce with similar conceivable consequences for the arrow of time as the above model that involves space inversion.

In cosmology, quantum aspects of the arrow of time must again play an important role. According to the Copenhagen interpretation, there is no quantum world – so no complete and consistent cosmic history would be defined any more when quantum aspects become essential. In other orthodox interpretations, the unitary evolution of the quantum state is repeatedly interrupted by measurements and similar time-asymmetric events, when the wave function is assumed to *collapse* indeterministically. The consequences of such stochastic events on quantum cosmology would be enormous, but as long as no general dynamical formulation of a collapse of the wave function has been confirmed, one has again arrived at an impasse. Going forward in time may be conceptually simple in such theories, since one just has to throw away all components of the wave function which represent the not actualized potential outcomes, while going backward would require these lost components to recombine and dynamically form local superpositions again. So one has to keep them in the cosmic bookkeeping at least – regardless of whether they are called *real* (as in the Everett interpretation) or not. Going back to the Big Bang would require *all* those many components that have ever been thrown away in the orthodox description during the past of our universe, while one would have to throw away others when formally going backward beyond the Big Bang in order to obtain an individual quasi-classical *pre-big-bang history*. In other words, a unitary continuation beyond the Big Bang can only relate the complete Everett superposition of worlds on both sides of the Big Bang, but *not* any individually observed quasi-classical worlds. The corresponding master arrow of time would thus not only affect all realms of physics – it must be truly universal in a much deeper sense: it can only have *multiversal* meaning. The same multiversality was required in a unitary black hole evolution of Sect. 10.3, and it does, in fact, apply to the unitary quantum description of all macroscopic objects, when irreversible decoherence mimics a collapse of the wave function and thereby explains classicality.

The time direction of Everett's branching of the wave function, based on decoherence, requires a homogeneous initial quantum state (presumably at $a = 0$), which does not contain any nonlocal entanglement that might later have local effects. Quantum dynamics will then lead to decoherence (the in practice irreversible dislocalization of superpositions), and thereby *intrinsically* break various global symmetries – possibly even in the form of many different quasi-classical *landscapes* that represent branches of one symmetric superposition of all of them.

10.5 Quantum Gravity

General Relativity has traditionally been considered in a block universe picture, but because of the hyperbolic type of Einstein's field equations, it is a dynamical theory just as any other field theory. The explicit dynamical description, which requires a non-Lorentz-invariant *form*, was completed by *Arnowitt*, *Deser*, and *Misner* (ADM) [10.32]. This Hamiltonian formulation is a prerequisite for the canonical quantization of the theory. I shall regard the result as an effective quantum theory, without considering any speculative generalizations or possible justifications (such as string theory or loop quantum gravity, respectively).

The ADM formalism is based on an arbitrary space-like foliation of spacetime that has to be chosen *on the fly*, that is, *while* solving an initial value problem. The spatial metric on these space-like slices represents the dynamical variables of the theory, and it has to be described by a symmetric matrix $h_{kl}(x_m)$ (with k, l, m running from 1 to 3). Three of its six independent matrix elements represent the choice of physically meaningless coordinates, two would in the linear limit correspond to the spin components of a gravitational wave (± 2 with respect to the direction of propagation for a plane wave), while the remaining one can be regarded as a measure of *many-fingered* physical time (metric distance between space-like slices). The corresponding canonical momenta π^{kl} define the embedding of the spatial metric into spacetime and the arbitrary propagation of spatial coordinates. The dynamics can then be formulated by means of the Hamiltonian equations with respect to an arbitrary time parameter t that formally distinguishes different slices in a given foliation. They are equivalent to Einstein's field equations. In contrast to metric time, the parameter t is geometrically or physically meaningless, and can therefore be replaced by any monotonic function $t' = f(t)$, including its inversion.

Note that when Special Relativity is said to abandon the concept of absolute time, this statement refers only to the concept of absolute simultaneity, while proper times, which control all motion according to the principle of relativity, are still assumed to be given *absolutely* by the fixed Lorentz metric. This remaining absoluteness is dropped only in General Relativity, where the metric itself becomes a dynamical object like matter, as described by the ADM formalism. The absence of an absolute time parameter (here represented by its reparametrizability) was already required by Ernst Mach. *Julian Barbour*, who studied its consequences in much historical detail [10.33–35], called it *timelessness*. However, a complete absence of time would remove any possibility of defining an arrow, while a remaining time parameter characterizing a one-dimensional succession of states still allows one to define time-asymmetric trajectories (histories).

The invariance of the theory under spatial coordinate transformations and time reparametrization is warranted by four constraints for the matrix $h_{kl}(t)$, called momentum and Hamiltonian constraints, respectively. They may be regarded as initial conditions, but are conserved in time. In particular, the Hamiltonian constraint assumes the form

$$H(h_{kl}, \pi_{kl}) = 0 . \tag{10.6}$$

When quantized (see [10.36] for a review), and when also taking into account matter variables, this constraint translates into the Wheeler–DeWitt equation,

$$H\Psi(h_{kl}, \text{matter}) = 0 , \tag{10.7}$$

which means that the time-dependent Schrödinger equation becomes trivial,

$$\frac{\partial \Psi}{\partial t} = 0 . \tag{10.8}$$

Even the time parameter t has now disappeared, because there are no trajectories in quantum theory any more that might be parametrized. Only this drastic quantum consequence of classical reparametrizability can really be regarded as *genuine timelessness*.

The timelessness of the Wheeler–DeWitt wave function has been known at least since 1967, but it seems to have originally been regarded as *just formal*. Time was often smuggled in again in various ways – for example in terms of parametrizable Feynman paths, by means of semiclassical approximations, or by attempts to reintroduce a Heisenberg picture in spite of the Hamiltonian constraint [10.37, 38]. The problem became pressing, though, in connection with realistic interpretations of the wave function in quantum cosmology [10.39–41].

The general wave functional Ψ (h_{kl}, matter) describes entanglement of geometry and matter. If we did have a succession of such quantum states (forming a quantum trajectory or quantum history), an appropriate, initially not entangled state could explain an arrow

of growing entanglement and decoherence. The resulting branching of the wave function according to a corresponding parameter t would then include branching spacetime geometries (that is, branching quasi-classical wave packets in the configuration space of three geometries). Although there is no time parameter any more, the metric h_{kl} still contains a measure of metric time. Therefore, it describes a *physical* time dependence in the form of an entanglement of this measure with all other degrees of freedom – even for a formally time-less solution of (10.7) [10.42]. For Friedmann universes, the expansion parameter a, which is part of the metric h_{kl}, is such an appropriate measure of time, but how does that help us to define an initial value problem for this static wave equation? The surprising answer is that this equation is globally hyperbolic for Friedmann-type universes – not on spacetime, as for classical fields, but on its infinite-dimensional configuration space (which has therefore also been called *superspace*). The expansion parameter a appears as a time-like variable in this sense because of the unusual negative sign of its formal kinetic energy component [10.43]. Therefore, the Wheeler–DeWitt equation allows one to define an *initial* value problem at a small value of a, for example. For a *modified* Wheeler–DeWitt equation, this possibility may even be extended to $a = 0$. There is no conceptual difference between a (multiversal) Big Bang and a Big Crunch any more, since in the absence of a time parameter the wave function can only be a standing wave on configuration space.

The metric tensor and other fields defined on a Friedmann sphere, $a = \text{const}$, may be represented by a four-dimensional multipole expansion, which is particularly useful for describing the very early, approximately homogeneous and isotropic universe [10.44, 45]. In this case, one may conveniently model matter quantum mechanically by a massive scalar field $\Phi(x_k)$. The wave functional of the universe then assumes the form $\Psi(a, \Phi_0, \{x_n\})$, where Φ_0 is the homogeneous part of the scalar field, while $\{x_n\}$ are all higher multipoles of geometry and matter. For the metric, only the tensor modes are geometrically meaningful, while the rest represents gauge degrees (here describing the propagation of spatial coordinates). The global hyperbolic nature with respect to all physical degrees of freedom becomes manifest in this representation.

In a simple toy model one neglects all higher multipoles in order to solve the Wheeler–DeWitt equation on the remaining two-dimensional *mini superspace* formed by the two monopoles only. The remaining Hamiltonian represents an a-dependent harmonic oscillator for Φ_0, which allows one to construct adiabatically stable Gaussian wave packets (*coherent states*) [10.46]. Figure 10.3 depicts the propagation of such a wave packet with respect to the *time* variable $\alpha = \ln a$. This standing wave on mini superspace mimics a timeless classical trajectory. However, the complete wave functional has to be expected to form a broad superposition of many such dynamically separate wave packets (a cosmologically early realization of *many worlds*). Note that these *worlds* are propagating wave packets rather than trajectories (as in David Deutsch's understanding of Everett). If the higher multipoles are also taken into account, the Wheeler–DeWitt equation may describe decoherence progressing with a – at first that of the monopoles Φ_0 and a itself, although this requires effective renormalization procedures [10.47].

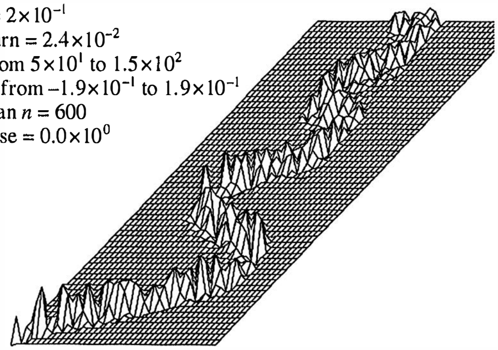


Fig. 10.3 Wave packet for a homogeneous massive scalar field amplitude Φ_0 (plotted *along the horizontal axis*) dynamically evolving as a function of the time-like parameter $\alpha = \ln a$ that is part of the metric (second axis in this two-dimensional mini superspace). The classical trajectory possesses a turning point above the plot region $50 \leq a \leq 150$ – namely at about $a = 240$ in this numerical example that represents an expanding and recontracting mini universe. Wave mechanically, this corresponds to a reflection of the wave packet by a repulsive potential in (10.5) at this value of a (reflected wave omitted in the plot). This reflection leads to considerable spreading of the *initial* wave packet. The causal order of these two legs of the trajectory is quite arbitrary, however, and the phase relations defining coherent wave packets could alternatively be chosen to give rise to a narrow wave packet for the second leg instead. So this (not shown) formal spreading does *not* represent a physical arrow of time (after [10.1, Sect. 6.2.1])

This intrinsic dynamics with respect to the time-like expansion parameter a has nothing as yet to do with the local dynamics in spacetime (controlled by proper times

along time-like curves) that must be relevant for matter as soon as the metric becomes quasi-classical. In order to understand the relation between these two kinds of dynamics, one may apply a Born–Oppenheimer expansion in terms of the inverse Planck mass, which is large compared to all particle masses, in order to study the Wheeler–DeWitt wave function [10.36, 48–50]. The Planck mass occurs in all kinetic energy terms of the geometric degrees of freedom that appear in the Hamiltonian constraint. The formal expansion in terms of powers of $m_{\text{Planck}}^{-1/4}$ then defines an *adiabatic approximation* with analogy to the theory of molecular motion (electron wave functions in the electrostatic fields of slowly moving nuclei). In most regions of configuration space (depending on the boundary conditions) one may further apply a Wentzel–Kramers–Brillouin *WKB* approximation to the *heavy* degrees of freedom Q. In this way one obtains an approximate solution of the type

$$\begin{aligned}\Psi(h_{kl}, \text{matter}) &= \Psi(Q, q)\\ &= e^{iS(Q)}\chi(Q, q)\,,\end{aligned} \tag{10.9}$$

where $S(Q)$ is a solution of the Hamilton–Jacobi equations for Q. The remaining wave function $\chi(Q, q)$ depends only slowly on Q, while q describes all light (matter) variables. Under these approximations one may derive from the Wheeler–DeWitt equation the adiabatic dependence of $\chi(Q, q)$ on Q in the form

$$i\nabla_Q S\nabla_Q\chi(Q, q) = h_Q\chi(Q, q)\,. \tag{10.10}$$

The operator h_Q is the weakly Q-dependent Hamiltonian for the matter variables q. This equation defines a new time parameter t_{WKB} separately along all WKB trajectories (which define classical spacetimes) by the directional derivative

$$\frac{\partial}{\partial t_{\text{WKB}}} := \nabla_Q S\nabla_Q\,. \tag{10.11}$$

In this way, one obtains from (10.10) a time-dependent global Schrödinger equation for matter with respect to the *derived* WKB time t_{WKB} [10.36, 39].This parameter defines a time coordinate in spacetime, since the classical trajectories $Q(t)$ in the superspace of spatial geometries Q define spacetime geometries. Equation (10.10) also describes the decoherence of superpositions of different WKB trajectories. Decoherence is also required to eliminate superpositions that form a real wave function $e^{iS}\chi + e^{-iS}\chi^*$, which has to be expected from the *real* Wheeler–DeWitt equation under physically meaningful boundary conditions.

In order to solve this derived time-dependent Schrödinger equation along a given WKB trajectory, that is, in terms of a foliation of a classical spacetime that does in turn adiabatically depend on the evolving matter, one needs a (low entropy) initial condition in the region where the WKB approximation begins to apply. For this purpose one would first have to solve the exact Wheeler–DeWitt equation (or its generalized version that may apply to some as yet elusive unified theory) as a function of a by using its fundamental cosmic initial condition at $a = 0$. This might be done, for example, by using the multipole expansion on the Friedmann sphere until one enters the WKB region (at some distance from $a = 0$), where this solution would then provide initial conditions for the matter wave functions χ for all arising WKB trajectories. The derived time-dependent Schrödinger equation with respect to t_{WKB} then has to be expected to describe further decoherence (the emergence of classical properties), and thereby explain the origin of all other arrows of time. In particular, it must enforce decoherence of superpositions of arising macroscopically different spacetimes, which would form separate quasi-classical *worlds* [10.36]. It would also decohere conceivable *CPT* symmetric superpositions of black and white holes, which are analogous to parity eigenstates of chiral molecules, if they had ever come into existence [10.23].

References

10.1 H.D. Zeh: *The Physical Basis of the Direction of Time*, 5th edn. (Springer, Berlin 2007)

10.2 J.R. Oppenheimer, H. Snyder: On continued gravitational contration, Phys. Rev. **56**, 455 (1939)

10.3 F.J. Dyson: Time without end: Physics and biology in an open universe, Rev. Mod. Phys. **51**, 447 (1979)

10.4 J.D. Bekenstein: Black holes and entropy, Phys. Rev. **D7**, 2333 (1973)

10.5 S.W. Hawking: Particle creation by black holes, Commun. Math. Phys. **43**, 199 (1975)

10.6 F.C. Adams, G. Laughlin: Aging universe: The long-term fate and evolution of astrophysical objects, Rev. Mod. Phys. **69**, 337 (1997)

10.7 D.N. Page: Is black hole evaporation predictable?, Phys. Lett. **B95**, 244 (1980)

10.8 Y.B. Zel'dovich: Gravity, charge, cosmology and coherence, Usp. Fiz. Nauk **123**, 487 (1977), Sov. Phys. Usp. **20**, 945 (1977)

10.9 S.W. Hawking: Breakdown of predictability in gravitational collapse, Phys. Rev. **D14**, 2460 (1976)

10.10 D.N. Page: Black hole information, Proc. 5th Can. Conf. Gen. Rel. Relat. Astrophys., ed. by R.B. Mann, R.G. McLenaghan (World Scientific, Singapore 1994), and reference therein
10.11 D. Gottesmann, J. Preskill: Comment on "The black hole final state", J. High Energy Phys. **0403**, 026 (2004)
10.12 S.W. Hawking: Information loss in black holes, Phys. Rev. **D72**, 084013 (2005)
10.13 S.D.H. Hsu, D. Reeb: Black holes, information and decoherence, Phys. Rev. **D79**, 124037 (2009)
10.14 C. Barceló, S. Liberati, S. Sonego, M. Visser: Analogue gravity, arXiv 1011.5911v1 (2010)
10.15 A. Almheiri, D. Marolf, J. Polchinski, J. Sully: Black holes: Complementarity of firewalls, arXiv:1207.3123v2 (2012)
10.16 C. Kiefer: Hawking radiation from decoherence, Class. Quantum Gravity **18**, L151 (2001)
10.17 C. Kiefer: Decoherence and entropy in complex systems. In: *Decoherence and Entropy in Complex Systems*, ed. by H.T. Elze (Springer, Berlin 2004)
10.18 H.D. Zeh: Where has all the information gone?, Phys. Lett. **A347**, 1 (2005)
10.19 R. Bousso: Firewalls from double purity, arxiv:1308.2665 (2013)
10.20 D. Page: Time dependence of Hawking radiation entropy, arxiv:1301.4995 (2013)
10.21 R. Penrose: Time-asymmetry and quantum gravity. In: *Quantum Gravity*, Vol. 2, ed. by C.J. Isham, R. Penrose, D.W. Sciama (Clarendon, London 1981) pp. 245–272
10.22 C. Kiefer: Can the arrow of time be understood from quantum gravity?, arXiv 0910.5836 (2009)
10.23 S.W. Hawking: Black holes and thermodynamics, Phys. Rev. **D13**, 191 (1976)
10.24 L. Boltzmann: Über irreversible Strahlungsvorgänge, Berl. Ber., pp. 1016–1018 (1897)
10.25 A. Einstein, W. Ritz: Zum gegenwärtigen Stand des Stahlungsproblems, Phys. Z. **10**, 323 (1911)
10.26 T. Gold: The arrow of time, Am. J. Phys. **30**, 403 (1962)
10.27 H.D. Zeh: Remarks on the compatiblity of opposite arrows of time, Entropy **8**, 44 (2006)
10.28 C. Kiefer, H.D. Zeh: Arrow of time in a recollapsing quantum universe, Phys. Rev. **D51**, 4145 (1995)
10.29 M. Bojowald: Initial conditions for a universe, Gen. Relativ. Gravit. **35**, 1877 (2003)
10.30 H.D. Conradi, H.D. Zeh: Quantum cosmology as an initial value problem, Phys. Lett. **A151**, 321 (1991)
10.31 A. Ashtekar, M. Campiglia, A. Henderson: Path integrals and WKB approximation in loop quantum cosmology, report arXiv 1011.1024v1 (2010)
10.32 R. Arnowitt, S. Deser, C.W. Misner: The dynamics of general relativity. In: *Gravitation: An Introduction to Current Research*, ed. by L. Witten (Wiley, New York 1962)
10.33 J. Barbour: Leibnizian time, Machian dynamics and quantum gravity. In: *Quantum Concepts in Space and Time*, ed. by R. Penrose, C.J. Isham (Cambridge Univ. Press, Cambridge 1986)
10.34 J. Barbour: The timelessnes of quantum gravity, Class. Quantum Gravity **11**, 2853 (1994)
10.35 J. Barbour: *The End of Time* (Weidenfeld and Nicolson, London 1999)
10.36 C. Kiefer: *Quantum Gravity* (Cambridge Univ. Press, Cambridge 2007)
10.37 K. Kuchar: Time and interpretations of quantum gravity, Proc. 4th Can. Conf. Gen. Rel. Relat. Astrophys., ed. by G. Kunstatter, D. Vincent, J. Williams (World Scientific, Singapore 1992)
10.38 C.J. Isham: Canonical quantum gravity and the problem of time. In: *Integrable Systems, Quantum Groups and Quantum Field Theory*, ed. by L.A. Ibort, M.A. Rodriguez (Kluwer, Doordrecht 1993)
10.39 H.D. Zeh: *Die Physik der Zeitrichtung*, Springer Lecture Notes (Springer, Berlin 1984), Chap. 6
10.40 H.D. Zeh: Emergence of classical time from a universe wave function, Phys. Lett. **A116**, 9 (1986)
10.41 H.D. Zeh: Time in quantum gravity, Phys. Lett. **A126**, 311 (1988)
10.42 D.N. Page, W.K. Wootters: Evolution without evolution: Dynamics described by stationary observables, Phys. Rev. **D27**, 2885 (1983)
10.43 D. Giulini, C. Kiefer: Wheeler–DeWitt metric and the attractivity of gravity, Phys. Lett. **A193**, 21 (1994)
10.44 J.J. Halliwell, S.W. Hawking: Origin of structure in the universe, Phys. Rev. **D31**, 1777 (1985)
10.45 C. Kiefer: Continuous measurement of minisuperspace variables by higher multipoles, Class. Quantum Gravity **4**, 1369 (1987)
10.46 C. Kiefer: Wave packets in minisuperspace, Phys. Rev. **D38**, 1761 (1988)
10.47 A.O. Barvinsky, A.Y. Kamenshchik, C. Kiefer, I.V. Mishakov: Decoherence in quantum cosmology at the onset of inflation, Nucl. Phys. **B551**, 374 (1999)
10.48 V.G. Lapchinsky, V.A. Rubakov: Canonical quantization of gravity and quantum field theory in curved space-time, Acta Phys. Pol. **10**, 1041 (1979)
10.49 T. Banks: TCP, quantum gravity, the cosmological constant and all that, Nucl. Phys. **B249**, 332 (1985)
10.50 R. Brout, G. Venturi: Time in semiclassical gravity, Phys. Rev. **D39**, 2436 (1989)

11. Teleparallelism: A New Insight into Gravity

José G. Pereira

Teleparallel gravity, a gauge theory for the translation group, turns up as fully equivalent to Einstein's general relativity. In spite of this equivalence, it provides a whole new insight into gravitation. It breaks several paradigms related to the geometric approach of general relativity, and introduces new concepts in the description of the gravitational interaction. The purpose of this chapter is to explore some of these concepts, as well as discuss possible consequences for gravitation, mainly those that could be relevant for the quantization of the gravitational field.

11.1 Preliminaries

Despite being equivalent to general relativity, teleparallel gravity is, conceptually speaking, a completely different theory. For example, the gravitational field in this theory is represented by torsion, not by curvature. Furthermore, in general relativity curvature is used to *geometrize* the gravitational interaction: geometry replaces the concept of gravitational force, and the trajectories are determined by geodesics – trajectories that follow the curvature of spacetime. Teleparallel gravity, on the other hand, attributes gravitation to torsion, which acts as a *force*, not geometry. In teleparallel gravity, therefore, trajectories are not described by geodesics, but by force equations [11.1].

The reason for gravitation to present two equivalent descriptions is related to its most peculiar property: *universality*. Like the other fundamental interactions of nature, gravitation can be described in terms of a gauge theory. This is just teleparallel gravity, a gauge theory for the translation group. Universality of free fall, on the other hand, allows a second, geometric description, based on the equivalence principle, just general relativity. As the unique universal interaction, it is the only one to allow a geometric interpretation, and hence two alternative descriptions. From this point of view, curvature and torsion are simply alternative ways of representing the very same gravitational field, accounting for the same degrees of freedom of gravity. (There are models in which curvature and torsion are related to different degrees of freedom of gravity. In these models, known as Einstein–Cartan–Sciama–Kibble theories, in addition to energy and momentum, also intrinsic spin appears as source of gravitation. The main references on these theories can be traced back from [11.2].)

The notion of teleparallel structure – also known as absolute or distant parallelism, characterized by a particular Lorentz connection that parallel-transports everywhere the tetrad field (Sect. 11.3.2 for a remark about the notion of absolute parallelism condition and local Lorentz transformations.) – was used by Einstein in his unsuccessful attempt to construct a unified field theory of electromagnetism and gravitation [11.3–5]. The birth of teleparallel gravity as a gravitational theory, however, took place in the late fifties and early sixties with the works by Møller [11.6]. Since then many contributions from different authors have been incorporated into the theory, giving rise to what is known today as the teleparallel equivalent of general relativity, or just teleparallel gravity [11.7]. The purpose of this chapter is to review the fundamentals of this theory, as well as to explore some of the new insights it provides into gravitation, in particular those that could eventually be relevant for the development of a quantum theory for gravitation.

11.2 Basic Concepts

11.2.1 Linear Frames and Tetrads

Spacetime is the arena on which the four presently known fundamental interactions manifest themselves. Electromagnetic, weak and strong interactions are described by gauge theories involving transformations taking place in *internal* spaces, by themselves unrelated to spacetime. The basic setting of gauge theories are the principal bundles, in which a copy of the gauge group is attached to each point of spacetime – the base space of the bundle. Gravitation, on the other hand, is deeply linked to the very structure of spacetime. The geometrical setting of gravitation is the tangent bundle, a natural construction always present in any differentiable manifold: at each point of spacetime there is a tangent space attached to it – the fiber of the bundle – which is seen as a vector space. We are going to use the Greek alphabet ($\mu, \nu, \rho, \ldots = 0, 1, 2, 3$) to denote indices related to spacetime, and the first letters of the Latin alphabet ($a, b, c, \ldots = 0, 1, 2, 3$) to denote indices related to the tangent space, a Minkowski spacetime whose Lorentz metric is assumed to have the form

$$\eta_{ab} = \text{diag}(+1, -1, -1, -1)\ . \tag{11.1}$$

A general spacetime is a 4-dimensional differential manifold, denoted $\mathbb{R}^{3,1}$, whose tangent space is, at each point, a Minkowski spacetime. Spacetime coordinates will be denoted by $\{x^\mu\}$, whereas tangent space coordinates will be denoted by $\{x^a\}$. Such coordinate systems determine, on their domains of definition, local bases for vector fields, formed by the sets of gradients

$$\{\partial_\mu\} \equiv \{\partial/\partial x^\mu\} \quad \text{and} \quad \{\partial_a\} \equiv \{\partial/\partial x^a\}\ , \tag{11.2}$$

as well as bases $\{dx^\mu\}$ and $\{dx^a\}$ for covector fields, or differentials. These bases are dual, in the sense that

$$dx^\mu(\partial_\nu) = \delta^\mu_\nu \quad \text{and} \quad dx^a(\partial_b) = \delta^a_b\ . \tag{11.3}$$

On the respective domains of definition, any vector or covector can be expressed in terms of these *coordinate bases*, a name that stems from their relationship to a coordinate system.

Trivial Frames

Trivial frames, or trivial tetrads [11.8], will be denoted by

$$\{e_a\} \quad \text{and} \quad \{e^a\}\ . \tag{11.4}$$

The above mentioned coordinate bases

$$\{e_a\} = \{\partial_a\} \quad \text{and} \quad \{e^a\} = \{dx^a\} \tag{11.5}$$

are very particular cases. Any other set of four linearly independent fields $\{e_a\}$ will form another basis, and will have a dual $\{e^a\}$ whose members are such that

$$e^a(e_b) = \delta^a_b\ . \tag{11.6}$$

Notice that, on a general manifold, vector fields are (like coordinate systems) only locally defined – and linear frames, as sets of four such fields, defined only on restricted domains.

These frame fields are the general linear bases on the spacetime differentiable manifold $\mathbb{R}^{3,1}$. The whole set of such bases, under conditions making of it also a differentiable manifold, constitutes the *bundle of linear frames*. A frame field provides, at each point $p \in \mathbb{R}^{3,1}$, a basis for the vectors on the tangent space $T_p\mathbb{R}^{3,1}$. Of course, on the common domains they are defined,

each member of a given basis can be written in terms of the members of any other. For example,

$$e_a = e_a^{\mu}\partial_\mu \quad \text{and} \quad e^a = e^a{}_\mu \, dx^\mu \,, \tag{11.7}$$

and conversely,

$$\partial_\mu = e^a{}_\mu e_a \quad \text{and} \quad dx^\mu = e_a^{\mu} e^a \,. \tag{11.8}$$

On account of the orthogonality conditions (11.6), the frame components satisfy

$$e^a{}_\mu e_a^{\nu} = \delta_\mu^\nu \quad \text{and} \quad e^a{}_\mu e_b^{\mu} = \delta_b^a \,. \tag{11.9}$$

Notice that these frames, with their bundles, are constitutive parts of spacetime: they are automatically present as soon as spacetime is taken to be a differentiable manifold.

A general linear basis $\{e_a\}$ satisfies the commutation relation

$$[e_a, e_b] = f^c{}_{ab}\, e_c \,, \tag{11.10}$$

with $f^c{}_{ab}$ the so-called structure coefficients, or coefficients of anholonomy, or still the anholonomy of frame $\{e_a\}$. As a simple computation shows, they are defined by

$$f^c{}_{ab} = e_a^{\mu} e_b^{\nu} (\partial_\nu e^c{}_\mu - \partial_\mu e^c{}_\nu). \tag{11.11}$$

A preferred class is that of inertial frames, denoted e'_a, those for which

$$f'^a{}_{cd} = 0 \,. \tag{11.12}$$

Such bases $\{e'^a\}$ are said to be *holonomic*. Of course, all coordinate bases are holonomic. This is not a local property, in the sense that it is valid everywhere for frames belonging to this inertial class.

Consider now the Minkowski spacetime metric, which in cartesian coordinates $\{\bar{x}^\mu\}$ has the form

$$\bar{\eta}_{\mu\nu} = \mathrm{diag}(+1, -1, -1, -1) \,. \tag{11.13}$$

In any other coordinate system, $\eta_{\mu\nu}$ will be a function of the spacetime coordinates. The linear frame

$$e_a = e_a^{\mu} \partial_\mu \,, \tag{11.14}$$

provides a relation between the tangent-space metric η_{ab} and the spacetime metric $\eta_{\mu\nu}$. Such relation is given by

$$\eta_{ab} = \eta_{\mu\nu} e_a^{\mu} e_b^{\nu} \,. \tag{11.15}$$

Using the orthogonality conditions (11.9), the inverse relation is found to be

$$\eta_{\mu\nu} = \eta_{ab} e^a{}_\mu e^b{}_\nu \,. \tag{11.16}$$

Independently of whether e_a is holonomic or not, or equivalently, whether they are inertial or not, they always relate the tangent Minkowski space to a Minkowski spacetime. These are the frames appearing in special relativity, and are usually called trivial frames – or trivial tetrads.

Nontrivial Frames

Nontrivial frames, or nontrivial tetrads, will be denoted by

$$\{h_a\} \quad \text{and} \quad \{h^a\} \,. \tag{11.17}$$

They are defined as linear frames whose coefficient of anholonomy is related to both inertial effects *and* gravitation. Let us consider a general pseudo-riemannian spacetime with metric components $g_{\mu\nu}$ in some dual holonomic basis $\{dx^\mu\}$. The tetrad field

$$h_a = h_a{}^\mu \partial_\mu \quad \text{and} \quad h^a = h^a{}_\mu \, dx^\mu \,, \tag{11.18}$$

is a linear basis that relates $g_{\mu\nu}$ to the tangent-space metric η_{ab} through the relation

$$\eta_{ab} = g_{\mu\nu} h_a{}^\mu h_b{}^\nu \,. \tag{11.19}$$

The components of the dual basis members $h^a = h^a{}_\nu \, dx^\nu$ satisfy

$$h^a{}_\mu h_a{}^\nu = \delta_\mu^\nu \quad \text{and} \quad h^a{}_\mu h_b{}^\mu = \delta_b^a \,, \tag{11.20}$$

so that (11.19) has the inverse

$$g_{\mu\nu} = \eta_{ab} h^a{}_\mu h^b{}_\nu \,. \tag{11.21}$$

It follows from these relations that

$$h \equiv \det(h^a{}_\mu) = \sqrt{-g} \,, \tag{11.22}$$

with $g = \det(g_{\mu\nu})$.

A tetrad basis $\{h_a\}$ satisfies the commutation relation

$$[h_a, h_b] = f^c{}_{ab} h_c \,, \tag{11.23}$$

with $f^c{}_{ab}$ the structure coefficients, or coefficients of anholonomy, of frame $\{h_a\}$. The basic difference in relation to the linear bases $\{e_a\}$ is that now the $f^c{}_{ab}$ represent both inertial effects and gravitation, and are given by

$$f^c{}_{ab} = h_a{}^\mu h_b{}^\nu \, (\partial_\nu h^c{}_\mu - \partial_\mu h^c{}_\nu) \,. \tag{11.24}$$

Although nontrivial tetrads are, by definition, anholonomic due to the presence of gravitation, it is still possible that *locally*, $f^c{}_{ab} = 0$. In this case, $\mathrm{d}h^a = 0$, which means that h^a is locally a closed differential form. In fact, if this holds at a point p, then there is a neighborhood around p on which functions (coordinates) x^a exist such that

$$h^a = \mathrm{d}x^a \ .$$

We say that a closed differential form is always locally integrable, or exact. This is the case of locally inertial frames, which are always holonomic. In these frames, inertial effects locally compensate for gravitation.

11.2.2 Lorentz Connections

A *Lorentz connection* A_μ, frequently referred to also as *spin connection*, is a 1-form assuming values in the Lie algebra of the Lorentz group,

$$A_\mu = \tfrac{1}{2} A^{ab}{}_\mu S_{ab} \,, \tag{11.25}$$

with S_{ab} a given representation of the Lorentz generators. As these generators are antisymmetric in the algebraic indices, $A^{ab}{}_\mu$ must be equally antisymmetric in order to be lorentzian. This connection defines the Fock–Ivanenko covariant derivative [11.9, 10]

$$\mathcal{D}_\mu = \partial_\mu - \frac{\mathrm{i}}{2} A^{ab}{}_\mu S_{ab} \,, \tag{11.26}$$

whose second part acts only on algebraic, tangent space indices. For a Lorentz vector field ϕ^c, for example, the representation of the Lorentz generators are matrices S_{ab} with entries [11.11]

$$(S_{ab})^c{}_d = \mathrm{i}\left(\eta_{bd}\delta^c_a - \eta_{ad}\delta^c_b\right) . \tag{11.27}$$

The Fock–Ivanenko derivative is, in this case,

$$\mathcal{D}_\mu \phi^c = \partial_\mu \phi^c + A^c{}_{d\mu}\phi^d \ . \tag{11.28}$$

On account of the soldered character of the tangent bundle, a tetrad field relates tangent space (or internal) tensors with spacetime (or external) tensors. For example, if ϕ^a is an internal, or Lorentz vector, then

$$\phi^\rho = h_a{}^\rho \phi^a \tag{11.29}$$

will be a spacetime vector. Conversely, we can write

$$\phi^a = h^a{}_\rho \phi^\rho \ . \tag{11.30}$$

On the other hand, due to its nontensorial character, a connection will acquire a vacuum, nonhomogeneous term, under the same operation,

$$\begin{aligned} \Gamma^\rho{}_{\nu\mu} &= h_a{}^\rho \partial_\mu h^a{}_\nu + h_a{}^\rho A^a{}_{b\mu} h^b{}_\nu \\ &\equiv h_a{}^\rho \mathcal{D}_\mu h^a{}_\nu \,, \end{aligned} \tag{11.31}$$

where $\mathcal{D}_\mu$ is the covariant derivative (11.28), in which the generators act on internal (or tangent space) indices only. The inverse relation is, consequently,

$$\begin{aligned} A^a{}_{b\mu} &= h^a{}_\rho \partial_\mu h_b{}^\rho + h^a{}_\rho \Gamma^\rho{}_{\nu\mu} h_b{}^\nu \\ &\equiv h^a{}_\rho \nabla_\mu h_b{}^\rho \,, \end{aligned} \tag{11.32}$$

where ∇_μ is the standard covariant derivative in the connection $\Gamma^\nu{}_{\rho\mu}$, which acts on external indices only. For a spacetime vector ϕ^ν, for example, it has the form

$$\nabla_\mu \phi^\nu = \partial_\mu \phi^\nu + \Gamma^\nu{}_{\rho\mu}\phi^\rho \ . \tag{11.33}$$

Using relations (11.29) and (11.30), it is easy to verify that [11.12]

$$\mathcal{D}_\mu \phi^d = h^d{}_\rho \nabla_\mu \phi^\rho \ . \tag{11.34}$$

Equations (11.31) and (11.32) are simply different ways of expressing the property that the total covariant derivative of the tetrad – that is, a covariant derivative with connection terms for both internal and external indices – vanishes identically

$$\partial_\mu h^a{}_\nu - \Gamma^\rho{}_{\nu\mu} h^a{}_\rho + A^a{}_{b\mu} h^b{}_\nu = 0 \ . \tag{11.35}$$

Behavior Under Lorentz Transformations

A local Lorentz transformation is a transformation of the tangent space coordinates x^a

$$x'^a = \Lambda^a{}_b(x) x^b \ . \tag{11.36}$$

Under such a transformation, the tetrad transforms according to

$$h'^a = \Lambda^a{}_b(x) h^b \ . \tag{11.37}$$

At each point of a riemannian spacetime, (11.21) only determines the tetrad up to transformations of the six-parameter Lorentz group in the tangent space indices. This means that there exists actually an infinity of tetrads $h_a{}^\mu$, each one relating the spacetime metric $g_{\mu\nu}$ to the tangent space metric η_{cd}. This means that any

other Lorentz-rotated tetrad $\{h'_a\}$ will also relate the same metrics

$$g_{\mu\nu} = \eta_{cd} h'^c{}_\mu h'^d{}_\nu \,. \tag{11.38}$$

Under a local Lorentz transformation $\Lambda^a{}_b(x)$, the spin connection undergoes the transformation

$$A'^a{}_{b\mu} = \Lambda^a{}_c(x) A^c{}_{d\mu} \Lambda_b{}^d(x) + \Lambda^a{}_c(x)\partial_\mu \Lambda_b{}^c(x) \,. \tag{11.39}$$

The last, nonhomogeneous term appears due to the nontensorial character of connections.

11.2.3 Curvature and Torsion

Curvature and torsion require a Lorentz connection to be defined [11.13]. Given a Lorentz connection $A^a{}_{b\mu}$, the corresponding curvature is a 2-form assuming values in the Lie algebra of the Lorentz group,

$$R_{\nu\mu} = \tfrac{1}{2} R^{ab}{}_{\nu\mu} S_{ab} \,. \tag{11.40}$$

Torsion is also a 2-form, but assuming values in the Lie algebra of the translation group,

$$T_{\nu\mu} = T^a{}_{\nu\mu} P_a \,, \tag{11.41}$$

with $P_a = \partial_a$ the translation generators. The curvature and torsion components are given, respectively, by

$$R^a{}_{b\nu\mu} = \partial_\nu A^a{}_{b\mu} - \partial_\mu A^a{}_{b\nu} + A^a{}_{e\nu} A^e{}_{b\mu} - A^a{}_{e\mu} A^e{}_{b\nu} \tag{11.42}$$

and

$$T^a{}_{\nu\mu} = \partial_\nu h^a{}_\mu - \partial_\mu h^a{}_\nu + A^a{}_{e\nu} h^e{}_\mu - A^a{}_{e\mu} h^e{}_\nu \,. \tag{11.43}$$

Through contraction with tetrads, these tensors can be written in spacetime-indexed forms

$$R^\rho{}_{\lambda\nu\mu} = h_a{}^\rho h^b{}_\lambda R^a{}_{b\nu\mu} \,, \tag{11.44}$$

and

$$T^\rho{}_{\nu\mu} = h_a{}^\rho T^a{}_{\nu\mu} \,. \tag{11.45}$$

Using relation (11.32), their components are found to be

$$R^\rho{}_{\lambda\nu\mu} = \partial_\nu \Gamma^\rho{}_{\lambda\mu} - \partial_\mu \Gamma^\rho{}_{\lambda\nu} + \Gamma^\rho{}_{\eta\nu} \Gamma^\eta{}_{\lambda\mu} - \Gamma^\rho{}_{\eta\mu} \Gamma^\eta{}_{\lambda\nu} \tag{11.46}$$

and

$$T^\rho{}_{\nu\mu} = \Gamma^\rho{}_{\mu\nu} - \Gamma^\rho{}_{\nu\mu} \,. \tag{11.47}$$

Since the spin connection $A^a{}_{b\nu}$ is a four-vector in the last index, it satisfies

$$A^a{}_{bc} = A^a{}_{b\nu} h_c{}^\nu \,. \tag{11.48}$$

It can thus be verified that, in the anholonomic basis $\{h_a\}$, the curvature and torsion components are given respectively by

$$R^a{}_{bcd} = h_c\,(A^a{}_{bd}) - h_d\,(A^a{}_{bc}) + A^a{}_{ec} A^e{}_{bd} - A^a{}_{ed} A^e{}_{bc} - f^e{}_{cd} A^a{}_{be} \tag{11.49}$$

and

$$T^a{}_{bc} = A^a{}_{cb} - A^a{}_{bc} - f^a{}_{bc} \,, \tag{11.50}$$

where, we recall, $h_c = h_c{}^\mu \partial_\mu$. Use of (11.50) for three different combinations of indices gives

$$A^a{}_{bc} = \tfrac{1}{2}\left(f_b{}^a{}_c + T_b{}^a{}_c + f_c{}^a{}_b + T_c{}^a{}_b - f^a{}_{bc} - T^a{}_{bc}\right) \,. \tag{11.51}$$

This expression can be rewritten in the form

$$A^a{}_{bc} = \overset{\circ}{A}{}^a{}_{bc} + K^a{}_{bc} \,, \tag{11.52}$$

where

$$\overset{\circ}{A}{}^a{}_{bc} = \tfrac{1}{2}\left(f_b{}^a{}_c + f_c{}^a{}_b - f^a{}_{bc}\right) \tag{11.53}$$

is the usual expression of the general relativity spin connection in terms of the coefficients of anholonomy, and

$$K^a{}_{bc} = \tfrac{1}{2}\left(T_b{}^a{}_c + T_c{}^a{}_b - T^a{}_{bc}\right) \tag{11.54}$$

is the contortion tensor. The corresponding expression in terms of the spacetime-indexed linear connection reads

$$\Gamma^\rho{}_{\mu\nu} = \overset{\circ}{\Gamma}{}^\rho{}_{\mu\nu} + K^\rho{}_{\mu\nu} \,, \tag{11.55}$$

where

$$\overset{\circ}{\Gamma}{}^\sigma{}_{\mu\nu} = \tfrac{1}{2} g^{\sigma\rho}\left(\partial_\mu g_{\rho\nu} + \partial_\nu g_{\rho\mu} - \partial_\rho g_{\mu\nu}\right) \tag{11.56}$$

is the zero-torsion Christoffel, or Levi-Civita connection, and

$$K^{\rho}{}_{\mu\nu} = \tfrac{1}{2}\left(T_{\nu}{}^{\rho}{}_{\mu} + T_{\mu}{}^{\rho}{}_{\nu} - T^{\rho}{}_{\mu\nu}\right) \tag{11.57}$$

is the spacetime-indexed contortion tensor. Equations (11.52) and (11.55) are actually the content of a theorem, which states that any Lorentz connection can be decomposed into the spin connection of general relativity plus the contortion tensor [11.13]. As is well-known, the Levi-Civita connection of a general spacetime metric has vanishing torsion, but nonvanishing curvature

$$\overset{\circ}{T}{}^{\rho}{}_{\nu\mu} = 0 \quad \text{and} \quad \overset{\circ}{R}{}^{\rho}{}_{\lambda\nu\mu} \neq 0\,. \tag{11.58}$$

11.2.4 Purely Inertial Lorentz Connection

In special relativity, Lorentz connections represent inertial effects present in a given frame. In order to obtain the explicit form of such connections, let us recall that the class of inertial (or holonomic) frames, denoted by $e'^{a}{}_{\mu}$, is defined by all frames for which $f'^{c}{}_{ab} = 0$. In a general coordinate system, the frames belonging to this class have the holonomic form

$$e'^{a}_{\mu} = \partial_{\mu} x'^{a}\,, \tag{11.59}$$

with x'^{a} a spacetime-dependent Lorentz vector: $x'^{a} = x'^{a}(x^{\mu})$. Under a local Lorentz transformation, the holonomic frame (11.59) transforms according to

$$e^{a}{}_{\mu} = \Lambda^{a}{}_{b}(x) e'^{b}{}_{\mu}\,. \tag{11.60}$$

As a simple computation shows, it has the explicit form

$$e^{a}{}_{\mu} = \partial_{\mu} x^{a} + \overset{\bullet}{A}{}^{a}{}_{b\mu} x^{b} \equiv \overset{\bullet}{\mathcal{D}}_{\mu} x^{a}\,, \tag{11.61}$$

where

$$\overset{\bullet}{A}{}^{a}{}_{b\mu} = \Lambda^{a}{}_{e}(x)\,\partial_{\mu}\Lambda_{b}{}^{e}(x) \tag{11.62}$$

is a Lorentz connection that represents the inertial effects present in the new frame $e^{a}{}_{\mu}$. As can be seen from (11.39), it is just the connection obtained from a Lorentz transformation of the vanishing spin connection $\overset{\bullet}{A}{}'^{e}{}_{d\mu} = 0$

$$\begin{aligned}\overset{\bullet}{A}{}^{a}{}_{b\mu} &= \Lambda^{a}{}_{e}(x)\overset{\bullet}{A}{}'^{e}{}_{d\mu}\Lambda_{b}{}^{d}(x)\\ &\quad + \Lambda^{a}{}_{e}(x)\,\partial_{\mu}\Lambda_{b}{}^{e}(x)\,.\end{aligned} \tag{11.63}$$

Starting from an inertial frame, different classes of frames are obtained by performing *local* (point-dependent) Lorentz transformations $\Lambda^{a}{}_{b}(x^{\mu})$. Within each class, the infinitely many frames are related through *global* (point-independent) Lorentz transformations, $\Lambda^{a}{}_{b} = $ constant.

Each component of the inertial connection (11.62), which is sometimes referred to as the Ricci coefficient of rotation [11.14], represents a different inertial effect [11.15]. Owing to its presence, the transformed frame $e^{a}{}_{\mu}$ is no longer holonomic. In fact, its coefficient of anholonomy is given by

$$f^{c}{}_{ab} = -\left(\overset{\bullet}{A}{}^{c}{}_{ab} - \overset{\bullet}{A}{}^{c}{}_{ba}\right), \tag{11.64}$$

where we have used the identity $\overset{\bullet}{A}{}^{a}{}_{bc} = \overset{\bullet}{A}{}^{a}{}_{b\mu} e_{c}{}^{\mu}$. Of course, as a purely inertial connection, $\overset{\bullet}{A}{}^{a}{}_{b\mu}$ has vanishing curvature and torsion

$$\begin{aligned}\overset{\bullet}{R}{}^{a}{}_{b\nu\mu} &\equiv \partial_{\nu}\overset{\bullet}{A}{}^{a}{}_{b\mu} - \partial_{\mu}\overset{\bullet}{A}{}^{a}{}_{b\nu} + \overset{\bullet}{A}{}^{a}{}_{e\nu}\overset{\bullet}{A}{}^{e}{}_{b\mu}\\ &\quad - \overset{\bullet}{A}{}^{a}{}_{e\mu}\overset{\bullet}{A}{}^{e}{}_{b\nu}\\ &= 0\end{aligned} \tag{11.65}$$

and

$$\begin{aligned}\overset{\bullet}{T}{}^{a}{}_{\nu\mu} &\equiv \partial_{\nu} e^{a}{}_{\mu} - \partial_{\mu} e^{a}{}_{\nu} + \overset{\bullet}{A}{}^{a}{}_{e\nu} e^{e}{}_{\mu}\\ &\quad - \overset{\bullet}{A}{}^{a}{}_{e\mu} e^{e}{}_{\nu}\\ &= 0\,.\end{aligned} \tag{11.66}$$

11.2.5 Equation of Motion of Free Particles

To see how a purely inertial connection shows up in a concrete example, let us consider the equation of motion of a free particle. In the class of inertial frames $e'^{a}{}_{\mu}$, such particle is described by the equation of motion

$$\frac{\mathrm{d}u'^{a}}{\mathrm{d}\sigma} = 0\,, \tag{11.67}$$

with u'^{a} the anholonomic four-velocity, and

$$\mathrm{d}\sigma^{2} = \eta_{\mu\nu}\,\mathrm{d}x^{\mu}\,\mathrm{d}x^{\nu} \tag{11.68}$$

the quadratic Minkowski interval. In an anholonomic frame $e^{a}{}_{\mu}$, related to $e'^{a}{}_{\mu}$ by the local Lorentz transformation (11.60), the equation of motion assumes

the manifestly covariant form under local Lorentz transformations

$$\frac{\mathrm{d}u^a}{\mathrm{d}\sigma} + \overset{\bullet}{A}{}^a{}_{b\mu} u^b u^\mu = 0\,, \tag{11.69}$$

where

$$u^a = \Lambda^a{}_b(x) u'^b \tag{11.70}$$

is the Lorentz transformed four-velocity, and

$$u^\mu = u^a e_a{}^\mu \tag{11.71}$$

is the usual, holonomic four-velocity

$$u^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}\sigma}\,. \tag{11.72}$$

Observe that the inertial forces coming from the frame noninertiality are represented by the inertial connection on the left-hand side of the equation (11.69), which is noncovariant by its very nature. Observe also that it is invariant under general coordinate transformations.

In terms of the holonomic four-velocity written in cartesian coordinates $\{\bar{x}^\mu\}$, the particle equation of motion has the form

$$\frac{\mathrm{d}\bar{u}^\rho}{\mathrm{d}\sigma} = 0\,. \tag{11.73}$$

Under a general coordinate transformation $\bar{x}^\mu \to x^\mu$, it assumes the manifestly covariant form under general coordinate transformations

$$\frac{\mathrm{d}u^\rho}{\mathrm{d}\sigma} + \overset{\bullet}{\gamma}{}^\rho{}_{\nu\mu} u^\nu u^\mu = 0\,, \tag{11.74}$$

where [11.16]

$$\overset{\bullet}{\gamma}{}^\rho{}_{\nu\mu} = \tfrac{1}{2}\eta^{\rho\lambda}(\partial_\nu \eta_{\lambda\mu} + \partial_\mu \eta_{\lambda\nu} - \partial_\lambda \eta_{\nu\mu}) \tag{11.75}$$

is a flat, coordinate-related connection, with $\eta_{\nu\mu}$ the Minkowski metric written in the general coordinate system $\{x^\mu\}$. Of course, since the equations of motion (11.69) and (11.74) describe the same free particle, they are equivalent ways of writing the same equation of motion. This means that connections (11.62) and (11.75) are different ways of writing the very same inertial connection. In fact, using relation (11.71), it is an easy task to verify that they are related by

$$\begin{aligned}\overset{\bullet}{A}{}^a{}_{b\mu} &= e^a{}_\rho \partial_\mu e_b{}^\rho + e^a{}_\rho \overset{\bullet}{\gamma}{}^\rho{}_{\nu\mu} e_b{}^\nu \\ &\equiv e^a{}_\rho \overset{\bullet}{\nabla}_\mu e_b{}^\rho\,,\end{aligned} \tag{11.76}$$

which is a relation of the form (11.32) between equivalent connections. We can then conclude that local Lorentz transformations are equivalent to general coordinate transformations in the sense that they give rise to the very same inertial connection. In Sect. 11.4.5 we will discuss further the implications of this equivalence for gravitation.

11.3 Teleparallel Gravity: A Brief Review

For the sake of completeness we present in this section a short review of teleparallel gravity, as well as discuss its equivalence to general relativity.

11.3.1 Translational Gauge Potential

Teleparallel gravity corresponds to a gauge theory for the translation group [11.7]. As such, the gravitational field is represented by a translational gauge potential $B^a{}_\mu$, a 1-form assuming values in the Lie algebra of the translation group,

$$B_\mu = B^a{}_\mu P_a\,, \tag{11.77}$$

with $P_a = \partial_a$ the translation generators. On account of the translational coupling prescription, it appears as the nontrivial part of the tetrad,

$$h^a{}_\mu = e^a{}_\mu + B^a{}_\mu\,, \tag{11.78}$$

where

$$e^a{}_\mu \equiv \overset{\bullet}{\mathcal{D}}_\mu x^a = \partial_\mu x^a + \overset{\bullet}{A}{}^a{}_{b\mu} x^b \tag{11.79}$$

is the trivial (nongravitational) tetrad (11.61). Under an infinitesimal gauge translation

$$\delta x^a = \varepsilon^b P_b x^a \equiv \varepsilon^a\,, \tag{11.80}$$

with $\varepsilon^a \equiv \varepsilon^a(x^\mu)$ the transformation parameters, the gravitational potential $B^a{}_\mu$ transforms according to

$$\delta B^a{}_\mu = -\overset{\bullet}{\mathcal{D}}_\mu \varepsilon^a\,. \tag{11.81}$$

The tetrad (11.78) is consequently gauge invariant

$$\delta h^a{}_\mu = 0 \,. \tag{11.82}$$

This is a matter of consistency as a gauge transformation cannot change the spacetime metric.

11.3.2 Teleparallel Spin Connection

The gravitational field in teleparallel gravity is fully represented by the translational gauge potential $B^a{}_\mu$. This means that in this theory Lorentz connections keep their special-relativistic role of representing inertial effects only. The fundamental Lorentz connection of teleparallel gravity is consequently the purely inertial connection (11.62), which has of course vanishing curvature

$$\begin{aligned} \dot{R}^a{}_{b\mu\nu} &= \partial_\mu \dot{A}^a{}_{b\nu} - \partial_\nu \dot{A}^a{}_{b\mu} + \dot{A}^a{}_{e\mu}\dot{A}^e{}_{b\nu} \\ &\quad - \dot{A}^a{}_{e\nu}\dot{A}^e{}_{b\mu} = 0 \,. \end{aligned} \tag{11.83}$$

For a tetrad involving a nontrivial translational gauge potential, that is, for

$$B^a{}_\mu \neq \dot{\mathcal{D}}_\mu \varepsilon^a \,, \tag{11.84}$$

its torsion will be nonvanishing

$$\begin{aligned} \dot{T}^a{}_{\mu\nu} &= \partial_\mu h^a{}_\nu - \partial_\nu h^a{}_\mu + \dot{A}^a{}_{e\mu} h^e{}_\nu \\ &\quad - \dot{A}^a{}_{e\nu} h^e{}_\mu \\ &\neq 0 \,. \end{aligned} \tag{11.85}$$

Using the trivial identity

$$\dot{\mathcal{D}}_\mu \dot{\mathcal{D}}_\nu x^a - \dot{\mathcal{D}}_\nu \dot{\mathcal{D}}_\mu x^a = 0 \,, \tag{11.86}$$

it can be rewritten in the form

$$\begin{aligned} \dot{T}^a{}_{\mu\nu} &= \partial_\mu B^a{}_\nu - \partial_\nu B^a{}_\mu + \dot{A}^a{}_{b\mu} B^b{}_\nu \\ &\quad - \dot{A}^a{}_{b\nu} B^b{}_\mu \,, \end{aligned} \tag{11.87}$$

which is the field strength of teleparallel gravity. In this theory, therefore, gravitation is represented by torsion, not by curvature. On account of the gauge invariance of the tetrad, the field strength is also invariant under gauge transformations

$$\dot{T}'^a{}_{\mu\nu} = \dot{T}^a{}_{\mu\nu} \,. \tag{11.88}$$

This is actually an expected result. In fact, considering that the generators of the adjoint representation are the coefficients of structure of the group taken as matrices, and considering that these coefficients vanish for abelian groups, fields belonging to the adjoint representation of abelian gauge theories will always be gauge invariant – a well-known property of electromagnetism.

The spacetime linear connection corresponding to the inertial spin connection (11.62) is

$$\begin{aligned} \dot{\Gamma}^\rho{}_{\nu\mu} &= h_a{}^\rho \partial_\mu h^a{}_\nu + h_a{}^\rho \dot{A}^a{}_{b\mu} h^b{}_\nu \\ &\equiv h_a{}^\rho \dot{\mathcal{D}}_\mu h^a{}_\nu \,. \end{aligned} \tag{11.89}$$

This is the so-called Weitzenböck connection. Its definition is equivalent to the identity

$$\partial_\mu h^a{}_\nu + \dot{A}^a{}_{b\mu} h^b{}_\nu - \dot{\Gamma}^\rho{}_{\nu\mu} h^a{}_\rho = 0 \,. \tag{11.90}$$

In the class of frames in which the spin connection $\dot{A}^a{}_{b\mu}$ vanishes, it reduces to

$$\partial_\mu h^a{}_\nu - \dot{\Gamma}^\rho{}_{\nu\mu} h^a{}_\rho = 0 \,, \tag{11.91}$$

which is the so-called absolute, or distant parallelism condition, from where teleparallel gravity got its name. It is important to remark that, at the time the term absolute, or distant parallelism condition was coined, no one was aware that this condition holds only on a very specific class of frames. The general expression valid in any frame is that given by (11.90). This means essentially the the tetrad is not actually parallel-transported everywhere by the Weitzenböck connection. The name *teleparallel gravity* is consequently not appropriate. Of course, for historical reasons we shall keep it.

11.3.3 Teleparallel Lagrangian

As a gauge theory for the translation group, the action functional of teleparallel gravity can be written in the form [11.17]

$$\dot{S} = \frac{1}{2ck} \int \eta_{ab} \dot{T}^a \wedge \star \dot{T}^b \,, \tag{11.92}$$

where

$$\dot{T}^a = \tfrac{1}{2} \dot{T}^a{}_{\mu\nu}\, dx^\mu \wedge dx^\nu \tag{11.93}$$

is the torsion 2-form, $\star\overset{\bullet}{T}{}^a$ is the corresponding dual form, and $k = 8\pi G/c^4$. More explicitly,

$$\overset{\bullet}{S} = \frac{1}{8ck}\int \eta_{ab}\overset{\bullet}{T}{}^a{}_{\mu\nu} \star \overset{\bullet}{T}{}^b{}_{\rho\sigma}\,\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu \wedge \mathrm{d}x^\rho \wedge \mathrm{d}x^\sigma\,. \tag{11.94}$$

Taking into account the identity

$$\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu \wedge \mathrm{d}x^\rho \wedge \mathrm{d}x^\sigma = -\epsilon^{\mu\nu\rho\sigma} h\,\mathrm{d}^4x\,, \tag{11.95}$$

with $h = \det(h^a{}_\mu)$, the action functional assumes the form

$$\overset{\bullet}{S} = -\frac{1}{8ck}\int \overset{\bullet}{T}_{a\mu\nu} \star \overset{\bullet}{T}{}^a{}_{\rho\sigma}\epsilon^{\mu\nu\rho\sigma}\, h\,\mathrm{d}^4x\,. \tag{11.96}$$

Using then the generalized dual definition for soldered bundles [11.18]

$$\star T^a{}_{\mu\nu} = \frac{h}{2}\epsilon_{\mu\nu\alpha\beta}S^{a\alpha\beta}\,, \tag{11.97}$$

it reduces to

$$\overset{\bullet}{S} = \frac{1}{4ck}\int \overset{\bullet}{T}{}^a{}_{\rho\sigma}\overset{\bullet}{S}{}_a{}^{\rho\sigma}\, h\,\mathrm{d}^4x\,, \tag{11.98}$$

where

$$\overset{\bullet}{S}{}_a{}^{\rho\sigma} \equiv -\overset{\bullet}{S}{}_a{}^{\sigma\rho} = h_a{}^\nu\left(\overset{\bullet}{K}{}^{\rho\sigma}{}_\nu - \delta_\nu{}^\sigma\, \overset{\bullet}{T}{}^{\theta\rho}{}_\theta + \delta_\nu{}^\rho\, \overset{\bullet}{T}{}^{\theta\sigma}{}_\theta\right) \tag{11.99}$$

is the superpotential, with

$$\overset{\bullet}{K}{}^{\rho\sigma}{}_\nu = \frac{1}{2}\left(\overset{\bullet}{T}{}^{\sigma\rho}{}_\nu + \overset{\bullet}{T}{}_\nu{}^{\rho\sigma} - \overset{\bullet}{T}{}^{\rho\sigma}{}_\nu\right) \tag{11.100}$$

the contortion of the teleparallel torsion. The lagrangian corresponding to the above action is [11.19]

$$\overset{\bullet}{\mathcal{L}} = \frac{h}{4k}\overset{\bullet}{T}_{a\mu\nu}\overset{\bullet}{S}{}^{a\mu\nu}\,. \tag{11.101}$$

Using relation (11.55) for the specific case of teleparallel torsion, it is possible to show that

$$\overset{\bullet}{\mathcal{L}} = \overset{\circ}{\mathcal{L}} - \partial_\mu\left(2hk^{-1}\overset{\bullet}{T}{}^{\nu\mu}{}_\nu\right), \tag{11.102}$$

where

$$\overset{\circ}{\mathcal{L}} = -\frac{\sqrt{-g}}{2k}\overset{\circ}{R} \tag{11.103}$$

is the Einstein–Hilbert lagrangian of general relativity. Up to a divergence, therefore, the teleparallel lagrangian is equivalent to the lagrangian of general relativity.

One may wonder why the lagrangians are equivalent up to a divergence term. To understand that, let us recall that the Einstein–Hilbert lagrangian (11.103) depends on the tetrad, as well as on its first and second derivatives. The terms containing second derivatives, however, reduce to a divergence term [11.20]. In consequence, it is possible to rewrite the Einstein–Hilbert lagrangian in a form stating this aspect explicitly,

$$\overset{\circ}{\mathcal{L}} = \overset{\circ}{\mathcal{L}}_1 + \partial_\mu(\sqrt{-g}w^\mu)\,, \tag{11.104}$$

where $\overset{\circ}{\mathcal{L}}_1$ is a lagrangian that depends solely on the tetrad and its first derivatives, and w^μ is a four-vector. On the other hand, the teleparallel lagrangian (11.101) depends only on the tetrad and its first derivative. The divergence term in the equivalence relation (11.102) is then necessary to account for the different orders of the teleparallel and the Einstein–Hilbert lagrangians. We mention in passing that in classical field theory the lagrangians involve only the field and its first derivative. We can then say that teleparallel gravity is more akin to a field theory than general relativity. In Sect. 11.4.4 this point will be discussed in further details.

11.3.4 Field Equations

Consider the lagrangian

$$\mathcal{L} = \overset{\bullet}{\mathcal{L}} + \mathcal{L}_s\,, \tag{11.105}$$

with $\mathcal{L}_s$ the lagrangian of a general source field. Variation with respect to the gauge potential $B^a{}_\rho$ (or equivalently, in terms of the tetrad $h^a{}_\rho$) yields the teleparallel version of the gravitational field equation

$$\partial_\sigma(h\overset{\bullet}{S}{}_a{}^{\rho\sigma}) - kh\overset{\bullet}{J}{}_a{}^\rho = kh\Theta_a{}^\rho\,. \tag{11.106}$$

In this equation,

$$\overset{\bullet}{hJ}_a{}^\rho \equiv -\frac{\partial \overset{\bullet}{\mathcal{L}}}{\partial h^a{}_\rho} = \frac{1}{k} h_a{}^\mu \overset{\bullet}{S}_c{}^{\nu\rho} \overset{\bullet}{T}{}^c{}_{\nu\mu} - \frac{h_a{}^\rho}{h} \overset{\bullet}{\mathcal{L}} + \frac{1}{k} \overset{\bullet}{A}{}^c{}_{a\sigma} \overset{\bullet}{S}_c{}^{\rho\sigma} \tag{11.107}$$

stands for the gauge current, which in this case represents the Noether energy-momentum pseudo-current of gravitation plus inertial effects [11.21], and

$$h\Theta_a{}^\rho = -\frac{\delta \mathcal{L}_s}{\delta h^a{}_\rho} \equiv -\left(\frac{\partial \mathcal{L}_s}{\partial h^a{}_\rho} - \partial_\mu \frac{\partial \mathcal{L}_s}{\partial_\mu \partial h^a{}_\rho}\right) \tag{11.108}$$

is the source energy-momentum tensor. Due to the antisymmetry of the superpotential in the last two indices, the total (gravitational plus inertial plus source) energy-momentum density is conserved in the ordinary sense

$$\partial_\rho \left(h\overset{\bullet}{J}_a{}^\rho + h\Theta_a{}^\rho\right) = 0 \,. \tag{11.109}$$

The left-hand side of the gravitational field equation (11.106) depends on $\overset{\bullet}{A}{}^a{}_{b\mu}$ only. Using identity (11.52) for the specific case of the inertial connection $\overset{\bullet}{A}{}^a{}_{b\mu}$,

$$\overset{\bullet}{A}{}^a{}_{b\mu} = \overset{\circ}{A}{}^a{}_{b\mu} + \overset{\bullet}{K}{}^a{}_{b\mu} \,, \tag{11.110}$$

through a lengthy but straightforward calculation, it can be rewritten in terms of $\overset{\circ}{A}{}^a{}_{b\mu}$ only

$$\partial_\sigma \left(h\overset{\bullet}{S}_a{}^{\rho\sigma}\right) - kh\overset{\bullet}{J}_a{}^\rho = h\left(\overset{\circ}{R}_a{}^\rho - \tfrac{1}{2} h_a{}^\rho \overset{\circ}{R}\right) . \tag{11.111}$$

As expected due to the equivalence between the corresponding lagrangians, the teleparallel field equation (11.106) is equivalent to Einstein's field equation

$$\overset{\circ}{R}_a{}^\rho - \tfrac{1}{2} h_a{}^\rho \overset{\circ}{R} = k\Theta_a{}^\rho \,. \tag{11.112}$$

Observe that the energy-momentum tensor appears as the source in both theories: as the source of curvature in general relativity, and as the source of torsion in teleparallel gravity. This is in agreement with the idea that curvature and torsion are related to the same degrees of freedom of the gravitational field.

11.4 Achievements of Teleparallel Gravity

Despite being equivalent to general relativity, teleparallel gravity shows many conceptual distinctive features. In this section we discuss some of these features, as well as explore their possible consequences for the study of both classical and quantum gravity.

11.4.1 Separating Inertial Effects from Gravitation

In teleparallel gravity, the tetrad field has the form

$$h^a{}_\mu = \overset{\bullet}{\mathcal{D}}_\mu x^a + B^a{}_\mu \,. \tag{11.113}$$

The first term on the right-hand side represents the frame and the inertial effects present on it. The second term, given by the translational gauge potential, represents gravitation only. This means that both inertia and gravitation are included in the tetrad $h^a{}_\mu$. As a consequence, its coefficient of anholonomy,

$$f^c{}_{ab} = h_a{}^\mu h_b{}^\nu (\partial_\nu h^c{}_\mu - \partial_\mu h^c{}_\nu) \,, \tag{11.114}$$

will also represent both inertia and gravitation. Of course, the same is true of the spin connection of general relativity,

$$\overset{\circ}{A}{}^a{}_{b\mu} = \tfrac{1}{2} h^c{}_\mu \left(f_b{}^a{}_c + f_c{}^a{}_b - f^a{}_{bc}\right) . \tag{11.115}$$

Now, according to the identity (11.110), such spin connection can be decomposed in the form

$$\overset{\circ}{A}{}^a{}_{b\mu} = \overset{\bullet}{A}{}^a{}_{b\mu} - \overset{\bullet}{K}{}^a{}_{b\mu} \,. \tag{11.116}$$

Since $\overset{\bullet}{A}{}^a{}_{b\mu}$ represents inertial effects only, whereas $\overset{\bullet}{K}{}^a{}_{b\mu}$ represents the gravitational field, the above identity amounts actually to a decomposition of the general

relativity spin connection (11.115) into inertial and gravitational parts.

To see that this is in fact the case, let us consider a locally inertial frame in which the spin connection of general relativity vanishes

$$\overset{\circ}{A}{}^{a}{}_{b\mu} \doteq 0 \,. \tag{11.117}$$

In such local frame, although present, gravitation becomes locally undetectable. Making use of identity (11.116), the local vanishing of $\overset{\circ}{A}{}^{a}{}_{b\mu}$ can be rewritten in the form

$$\overset{\bullet}{A}{}^{a}{}_{b\mu} \doteq \overset{\bullet}{K}{}^{a}{}_{b\mu} \,. \tag{11.118}$$

This expression shows explicitly that, in such a local frame inertial effects (left-hand side) exactly compensate for gravitation (right-hand side) [11.22]. The possibility of separating inertial effects from gravitation is an outstanding property of teleparallel gravity. It opens up many interesting new roads for the study of gravitation, which are not possible in the context of general relativity.

11.4.2 Geometry Versus Force

In general relativity, the trajectories of spinless particles are described by the geodesic equation

$$\frac{\mathrm{d}u^a}{\mathrm{d}s} + \overset{\circ}{A}{}^{a}{}_{b\mu} u^b u^\mu = 0 \,, \tag{11.119}$$

where $\mathrm{d}s^2 = g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu$ is the riemannian spacetime quadratic interval. It says essentially that the four-acceleration of the particle vanishes

$$\overset{\circ}{a}{}^{a} = 0 \,. \tag{11.120}$$

This means that in general relativity *there is no the concept of gravitational force*. Using identity (11.116), the geodesic equation can be rewritten in terms of a purely inertial connection and its torsion. The result is

$$\frac{\mathrm{d}u^a}{\mathrm{d}s} + \overset{\bullet}{A}{}^{a}{}_{b\mu} u^b u^\mu = \overset{\bullet}{K}{}^{a}{}_{b\mu} u^b u^\mu \,. \tag{11.121}$$

This is the teleparallel equation of motion of a spinless particle as seen from a general Lorentz frame. Of course, it is equivalent to the geodesic equation (11.119). There are conceptual differences, though. In general relativity, a theory fundamentally based on the equivalence principle, curvature is used to *geometrize* the gravitational interaction. The gravitational interaction in this case is described by letting (spinless) particles to follow the curvature of spacetime. Geometry replaces the concept of force, and the trajectories are determined, not by force equations, but by geodesics. Teleparallel gravity, on the other hand, attributes gravitation to torsion, which accounts for gravitation not by geometrizing the interaction, but by acting as a force [11.1]. In consequence, there are no geodesics in teleparallel gravity, only force equations similar to the Lorentz force equation of electrodynamics. (We remark in passing that this is in agreement with the gauge structure of teleparallel gravity in the sense that gauge theories always describe the (classical) interaction through a force.) Notice that the inertial forces coming from the frame noninertiality are represented by the connection on the left-hand side, which is noncovariant by its very nature. In teleparallel gravity, therefore, whereas the gravitational effects are described by a covariant force, the inertial effects of the frame remain *geometrized* in the sense of general relativity. In the geodesic equation (11.119), both inertial and gravitational effects are described by the connection term on the left-hand side.

11.4.3 Gravitational Energy-Momentum Density

All fundamental fields have a well-defined local energy-momentum density. It is then expected that the same should happen to the gravitational field. However, no tensorial expression for the gravitational energy-momentum density can be defined in the context of general relativity. The basic reason for this impossibility is that both gravitational and inertial effects are mixed in the spin connection of the theory, and cannot be separated. Even though some quantities, like curvature, are not affected by inertial effects, some others turn out to depend on it. For example, the energy-momentum density of gravitation will necessarily include both the energy-momentum density of gravity and the energy-momentum density of the inertial effects present in the frame. Since the inertial effects are essentially nontensorial – they depend on the frame – the quantity defining the energy-momentum density of the gravitational field in this theory always shows up as a nontensorial object. Some examples of different pseudotensors can be found in [11.23–31].

On the other hand, owing to the possibility of separating gravitation from inertial effects in teleparallel gravity, it turns out possible to write down an energy-momentum density for gravitation only, excluding the contribution from inertia. Such quantity is a tensorial object. To see how this is possible, let us consider the sourceless version of the teleparallel field equation (11.106),

$$\partial_\sigma \left(h \overset{\bullet}{S}_a{}^{\rho\sigma} \right) - k h \overset{\bullet}{J}_a{}^\rho = 0 \,, \tag{11.122}$$

where

$$\begin{aligned} h \overset{\bullet}{J}_a{}^\rho = & \frac{1}{k} h_a{}^\mu \overset{\bullet}{S}_c{}^{\nu\rho} \overset{\bullet}{T}{}^c{}_{\nu\mu} \\ & - \frac{h_a{}^\rho}{h} \overset{\bullet}{\mathcal{L}} + \frac{1}{k} \overset{\bullet}{A}{}^c{}_{a\sigma} \overset{\bullet}{S}_c{}^{\rho\sigma} \end{aligned} \tag{11.123}$$

is the usual gravitational energy-momentum pseudo-current, which is conserved in the ordinary sense

$$\partial_\rho \left(h \overset{\bullet}{J}_a{}^\rho \right) = 0 \,. \tag{11.124}$$

This is actually a matter of necessity: since the derivative is not covariant, the conserved current cannot be covariant either so that the conservation law itself is covariant – and consequently physically meaningful.

Using now the fact that the last, nontensorial term of the pseudo-current (11.123) together with the potential term make up a Fock–Ivanenko covariant derivative,

$$\partial_\sigma \left(h \overset{\bullet}{S}_a{}^{\rho\sigma} \right) - \overset{\bullet}{A}{}^c{}_{a\sigma} \left(h \overset{\bullet}{S}_c{}^{\rho\sigma} \right) \equiv \overset{\bullet}{\mathcal{D}}_\sigma \left(h \overset{\bullet}{S}_a{}^{\rho\sigma} \right) , \tag{11.125}$$

the field equation (11.122) can be rewritten in the form

$$\overset{\bullet}{\mathcal{D}}_\sigma \left(h \overset{\bullet}{S}_a{}^{\rho\sigma} \right) - k h \overset{\bullet}{t}_a{}^\rho = 0 \,, \tag{11.126}$$

where

$$\overset{\bullet}{t}_a{}^\rho = \frac{1}{k} h_a{}^\lambda \overset{\bullet}{S}_c{}^{\nu\rho} \overset{\bullet}{T}{}^c{}_{\nu\lambda} - \frac{h_a{}^\rho}{h} \overset{\bullet}{\mathcal{L}} \tag{11.127}$$

is a tensorial current that represents the energy-momentum of gravity alone [11.21]. Considering that the teleparallel spin connection (11.62) has vanishing curvature, the corresponding Fock–Ivanenko derivative is commutative

$$\left[\overset{\bullet}{\mathcal{D}}_\rho , \overset{\bullet}{\mathcal{D}}_\sigma \right] = 0 \,. \tag{11.128}$$

Taking into account the anti-symmetry of the superpotential in the last two indices, it follows from the field equation (11.126) that the tensorial current (11.127) is conserved in the covariant sense

$$\overset{\bullet}{\mathcal{D}}_\rho \left(h \overset{\bullet}{t}_a{}^\rho \right) = 0 \,. \tag{11.129}$$

This is again a matter of necessity: a covariant current can only be conserved in the covariant sense. Of course, since it does not represent the total energy-momentum density – in the sense that the inertial energy-momentum density is not included – it does not need to be truly conserved. Only the total energy-momentum density $\overset{\bullet}{J}_a{}^\rho$ must be truly conserved.

It should be remarked that the use of pseudotensors to compute the energy of a gravitational system requires some amount of handwork to get the physically relevant result. The reason is that, since the pseudotensor includes the contribution from the inertial effects, which is in general divergent for large distances (recall the centrifugal force, for example), the space integration of the energy density usually yields divergent results. It is then necessary to use appropriate coordinates – like for example cartesian coordinates [11.32] – or to make use of a regularization process to eliminate the spurious contributions coming from the inertial effects [11.33]. On the other hand, on account of the tensorial character of the teleparallel energy-momentum density of gravity, its use to compute the energy of any gravitational system always gives the physical result, no matter the coordinates or frames used to make the computation, eliminating in this way the necessity of using appropriate coordinates or a regularizing process [11.34].

11.4.4 A Genuine Gravitational Variable

Due to the fact that the spin connection of general relativity involves both gravitation and inertial effects, it is always possible to find a local frame – called *locally inertial frame* – in which inertial effects exactly compensate for gravitation, and the connection vanishes at that point

$$\overset{\circ}{A}{}^a{}_{b\mu} \doteq 0 \,. \tag{11.130}$$

Since there is gravitational field at that point, such connection cannot be considered a genuine gravitational variable in the usual sense of classical field theory. Strictly speaking, therefore, general relativity is not a true field theory in the usual sense of classical field theory. There is an additional problem: the noncovariant behavior of $\overset{\circ}{A}{}^a{}_{b\mu}$ under local Lorentz transformations is due uniquely to its inertial content, not to gravitation itself. To see it, consider the decomposition (11.116): whereas the first term on the right-hand side represents its inertial, noncovariant part, the second term represents its gravitational part, which is a tensor. This means that it is not a genuine gravitational connection either, but an inertial connection.

In teleparallel gravity, on the other hand, the gravitational field is represented by a translational-valued gauge potential

$$B_\mu = B^a{}_\mu P_a \,, \tag{11.131}$$

which shows up as the nontrivial part of the tetrad. Considering that the translational gauge potential represents gravitation only, not inertial effects, it can be considered a true gravitational variable in the sense of classical field theory. Notice, for example, that it is not possible to find a local frame in which it vanishes at a point. Furthermore, it is also a genuine gravitational connection: its connection behavior under gauge translations is related uniquely to its gravitational content. Put together, these properties show that, in contrast to general relativity, teleparallel gravity is a (background-dependent) true field theory.

11.4.5 Gravitation and Gauge Theories

If general relativity is not a true field theory, it cannot be a gauge theory either. There have been some attempts to describe general relativity as a gauge theory for diffeomorphisms, but this is impossible for several reasons. To begin with, general coordinate transformations take place on spacetime, not on the tangent space – the fiber of the tangent bundle – as it should be for a true gauge theory. In addition, general covariance by itself is empty of dynamical content in the sense that any relativistic equation, like for example Maxwell equation, can be written in a generally covariant form without any gravitational implication. There have also been some attempts to recast general relativity as a gauge theory for the Lorentz group. However, this is not possible either for different reasons. First, the spin connection of general relativity, as discussed in the previous section, is neither a true field variable nor a genuine gravitational connection. A second reason is that local Lorentz transformations are equivalent to general coordinate transformations in the sense that they give rise to the very same inertial connection.

Indeed, observe that the inertial connection (11.62), obtained by performing a *local Lorentz transformation*, and the inertial connection (11.75), obtained by performing a *general coordinate transformation*, represent two different ways of expressing the very same inertial connection, as shown by (11.76). Consciously or not, this equivalence is implicitly assumed in the metric formulation of general relativity. For example, it is a commonplace in many textbooks on gravitation to find the definition of a *locally inertial coordinate system*. Of course, the property of being or not inertial belongs to frames, not to coordinate systems. Such notion only makes sense if local Lorentz transformations and general coordinate transformations are considered on an equal footing. Then comes the point: since diffeomorphism is empty of dynamical meaning, and considering that it is equivalent to a local Lorentz transformation, the latter is also empty of dynamical meaning. One should not expect, therefore, any dynamical effect coming from a *gaugefication* of the Lorentz group.

On the other hand, there is a consistent rationale behind a gauge theory for the translation group. To begin with, remember that the source of gravitation is energy and momentum. From Noether's theorem, a fundamental piece of gauge theories [11.35], we know that the energy-momentum tensor is conserved provided the source lagrangian is invariant under spacetime translations. If gravity is to be described by a gauge theory with energy-momentum as source, therefore, it must be a gauge theory for the translation group. This is similar to electrodynamics, whose source lagrangian is invariant under the one-dimensional unitary group $U(1)$, the gauge group of Maxwell theory.

11.4.6 Gravity and the Quantum

If general relativity is not a field theory in the usual sense of the term, the traditional approach of quantum field theory cannot be used in this case. In addition, due to the fact that general relativity is deeply rooted on the equivalence principle, its spin connection involves both gravitation and inertial effects. As a consequence, any approach to quantum gravity using this connection as field variable will necessarily include a quantization of the inertial forces – whatever this may come to mean. Considering furthermore the di-

vergent asymptotic behavior of inertial effects, like for example the centrifugal force, such approach is likely to face consistency problems. As a matter of fact, in the geometric approach of general relativity there is not a genuine gravitational variable to be quantized using the methods of quantum field theory. For these reasons, one should not expect to obtain a consistent quantum gravity theory from general relativity. (Different arguments leading to the same conclusion can be found in [11.36].)

On the other hand, as a gauge theory for the translation group, teleparallel gravity is much more akin to a classical field theory than general relativity. It is, of course, different from the Yang–Mills type theories because of the soldering, which makes it a background-dependent field theory. In this theory, whereas inertial effects are represented by a Lorentz connection, the gravitational field is represented by a translational-valued connection, a legitimate gravitational variable in the usual sense of classical field theory. It is, for this reason, the variable to be quantized in any approach to quantum gravity. Taking into account that loop quantum gravity has a natural affinity with gauge theories [11.37–39], a quantization approach based on teleparallel gravity seems to be more consistent – and of course much simpler due to the abelian character of translations.

Still in connection to a prospective quantum theory for gravitation, it is important to remark that, differently from the geometrical approach of general relativity, the gauge approach of teleparallel gravity is not grounded on the equivalence principle [11.40]. In other words, it does not make use of the local equivalence between gravitation and inertial effects. As a consequence, it does not make use of ideal, local observers, as required by the strong equivalence principle, eliminating in this way the basic inconsistency with quantum mechanics, which presupposes real, dimensional observers [11.41]. Of course, this is not enough to guarantee that a quantum version of teleparallel gravity will be a consistent theory, but can be considered an important conceptual advantage of teleparallel gravity.

11.5 Final Remarks

Although equivalent to general relativity, teleparallel gravity introduces new concepts into both classical and quantum gravity. For example, on account of the geometric description of general relativity, which makes use of the torsionless Levi-Civita connection, there is a widespread belief that gravity produces a curvature in spacetime. The universe as a whole, in consequence, should also be curved. However, the advent of teleparallel gravity breaks this paradigm: it becomes a matter of convention to describe the gravitational interaction in terms of curvature or in terms of torsion. This means that the attribution of curvature to spacetime is not an absolute, but a model-dependent statement. Notice furthermore that, according to teleparallel gravity, torsion has already been detected: it is responsible for all gravitational phenomena, including the physics of the solar system, which can be re-interpreted in terms of a force equation with torsion playing the role of force. A reappraisal of cosmology based on teleparallel gravity could provide a new way to look at the universe, eventually unveiling new perspectives not visible in the standard approach based on general relativity.

Not only cosmology, but many other gravitational phenomena would acquire a new perspective when analyzed from the teleparallel point of view. For instance, in teleparallel gravity there is a tensorial expression for the energy-momentum density of gravitation alone, to the exclusion of inertial effects. Gravitational waves would no longer be interpreted as the propagation of curvature-perturbation in the fabric of spacetime, but as the propagation of torsional field-strength waves. Furthermore, similarly to the teleparallel gauge potential, a fundamental spin-2 field should be interpreted, not as a symmetric second rank tensor, but as a translational-valued vector field [11.42]. Most importantly, teleparallel gravity seems to be a much more appropriate theory to deal with the quantization of the gravitational field. We can then say that this theory is not just equivalent to general relativity, but a new way to look at all gravitational phenomena.

References

11.1 V.C. de Andrade, J.G. Pereira: Gravitational Lorentz force and the description of the gravitational interaction, Phys. Rev. D **56**, 4689 (1997), arXiv:gr-qc/9703059

11.2 M. Blagojević, F.W. Hehl (Eds.): *Gauge Theories of Gravitation. A Reader with Commentaries* (World Scientific/Imperial College Press, London 2012)

11.3 A. Einstein: Auf die Riemann-Metrik und den Fern-Parallelismus gegründete einheitliche Feldtheorie, Math. Annal. **102**, 685 (1930)

11.4 A. Unzicker, T. Case: Unified field theory based on Riemannian metrics and distant parallelism (2005) arXiv:physics/0503046

11.5 T. Sauer: Field equations in teleparallel spacetime: Einstein's 'Fernparallelismus' approach towards unified field theory (Einstein's Papers Project 2004) arXiv:physics/0405142

11.6 C. Møller: Conservation laws and absolute parallelism in general relativity, K. Dan. Vidensk. Selsk. Mat. Fys. Skr. **1**(10), 1–50 (1961)

11.7 R. Aldrovandi, J.G. Pereira: *Teleparallel Gravity: An Introduction* (Springer, Dordrecht 2012)

11.8 R. Aldrovandi, J.G. Pereira: *An Introduction to Geometrical Physics* (World Scientific, Singapore 1995)

11.9 V.A. Fock, D. Ivanenko: Über eine Mögliche Geometrische Deutung der Relativistischen Quantentheorie, Z. Phys. **54**, 798 (1929)

11.10 V.A. Fock: Geometrisierung der Diracschen Theorie des Elektrons, Z. Phys. **57**, 261 (1929)

11.11 P. Ramond: *Field Theory: A Modern Primer*, 2nd edn. (Addison-Wesley, Redwood 1989)

11.12 T.W.B. Kibble: Lorentz invariance and the gravitational field, J. Math. Phys. **2**, 212 (1961)

11.13 S. Kobayashi, K. Nomizu: *Foundations of Differential Geometry*, 2nd edn. (Wiley-Interscience, New York 1996)

11.14 C.W. Misner, K.S. Thorne, J.A. Wheeler: *Gravitation* (Freeman, San Francisco 1973)

11.15 D. Singh, G. Papini: Spin-1/2 particles in noninertial reference frames: Low- and high-energy approximations, Nuovo Cim. B **115**, 223 (2000), arXiv:gr-qc/0007032

11.16 R.A. Mosna, J.G. Pereira: Some remarks on the coupling prescription of teleparallel gravity, Gen. Relativ. Gravit. **36**, 2525 (2004), arXiv:gr-qc/0312093

11.17 L.D. Faddeev, A.A. Slavnov: *Gauge Fields* (Benjamin/Cummings, Reading 1980)

11.18 T.L. Gribl, J.G. Pereira: Hodge dual for soldered bundles, J. Phys. A **42**, 035402 (2009), arXiv:0811.2066

11.19 J.W. Maluf: Hamiltonian formulation of the teleparallel description of general relativity, J. Math. Phys. **35**, 335 (1994)

11.20 L.D. Landau, E.M. Lifshitz: *The Classical Theory of Fields* (Pergamon, Oxford 1975)

11.21 V.C. de Andrade, L.C.T. Guillen, J.G. Pereira: Gravitational energy-momentum density in teleparallel gravity, Phys. Rev. Lett. **84**, 4533 (2000), arXiv:gr-qc/0003100

11.22 R. Aldrovandi, L.C.T. Guillen, J.G. Pereira, K.H. Vu: Bringing together gravity and the quanta, Albert Einstein Century Int. Conf., Vol. 861, ed. by J.-M. Alimi, A. Füzfa (American Institute of Physics, New York 2006)

11.23 R.C. Tolman: *Relativity, Thermodynamics and Cosmology* (Oxford Univ. Press, Oxford 1934)

11.24 A. Papapetrou: Spinning test particles in general relativity, Proc. R. Soc. A **64**, 248 (1952)

11.25 P.G. Bergmann, R. Thompson: Spin and angular momentum in general relativity, Phys. Rev. **89**, 400 (1953)

11.26 C. Møller: On the localization of the energy of a physical system in the general theory of relativity, Ann. Phys. (NY) **4**, 347 (1958)

11.27 L.B. Szabados: On Canonical Pseudotensors, Sparling's form and noether currents, Class. Quantum Gravity **9**, 2521 (1992)

11.28 J.M. Aguirregabiria, A. Chamorro, K.S. Virbhadra: Energy and angular momentum of charged rotating black holes, Gen. Relativ. Gravit. **28**, 1393 (1996), arXiv:gr-qc/9501002

11.29 J.W. Maluf: Sparling two-forms, the conformal factor and the gravitational energy density of the teleparallel equivalent of general relativity, Gen. Relativ. Gravit. **30**, 413 (1998), arXiv:gr-qc/9710124

11.30 S. Deser, J.S. Franklin, D. Seminara: Graviton-graviton scattering, Bel-Robinson and energy (pseudo)-tensors, Class. Quantum Gravity **16**, 2815 (1999), arXiv:gr-qc/9905021

11.31 S.V. Babak, L.P. Grishchuk: Energy-momentum tensor for the gravitational field, Phys. Rev. D **61**, 024038 (2000), arXiv:gr-qc/9907027

11.32 N. Rosen, K.S. Virbhadra: Energy and momentum of cylindrical gravitational waves, Gen. Relativ. Gravit. **25**, 429 (1993)

11.33 J.W. Maluf, M.V.O. Veiga, J.F. da Rocha-Neto: Regularized expression for the gravitational energy-momentum in teleparallel gravity and the principle of equivalence, Gen. Relativ. Gravit. **39**, 227 (2007), arXiv:gr-qc/0507122

11.34 T.L. Gribl, N.Y. Obukhov, J.G. Pereira: Regularizing role of teleparallelism, Phys. Rev. D **80**, 064043 (2009), arXiv:0909.2418

11.35 N.P. Konopleva, V.N. Popov: *Gauge Fields* (Harwood, New York 1980)

11.36 V. Petkov: *Inertia and Gravitation: From Aristotle's Natural Motion to Geodesic Worldlines in Curved Spacetime* (Minkowski Institute Press, Montreal 2012), Chap. 6 and Appendix C

11.37 R. Gambini, J. Pullin: *Loops, Knots, Gauge Theories and Quantum Gravity* (Cambridge Univ. Press, Cambridge 1996)

11.38 C. Rovelli: *Quantum Gravity* (Cambridge Univ. Press, Cambridge 2004)

11.39 T. Thiemann: Lectures on loop quantum gravity, LNP **631**, 41 (2003), arXiv:gr-qc/0210094

11.40 R. Aldrovandi, J.G. Pereira, K.H. Vu: Gravitation without the equivalence principle, Gen. Relativ. Gravit. **36**, 101 (2004), arXiv:gr-qc/0304106

11.41 R. Aldrovandi, J.P. Beltrán Almeida, C.S.O. Mayor, J.G. Pereira: de Sitter Relativity and Quantum Physics, Quantum Theory: Reconsideration of Foundations Vol. 4, Vol. 962, ed. by G. Adenier, A. Khrennikov, P. Lahti, V.I. Man'ko, T. Nieuwenhuizen (2007) p. 175, arXiv:0710.0610

11.42 H.I. Arcos, T.L. Gribl, J.G. Pereira: Consistent gravitationally-coupled spin-2 field theory, Class. Quantum Gravity **27**, 145007 (2010), arXiv:1001.3407

12. Gravity and Spacetime: An Emergent Perspective

Thanu Padmanabhan

Classical general relativity treats spacetime as a continuum just as fluid dynamics treats a fluid as a continuum. Boltzmann was the first to emphasize that the thermal phenomena exhibited by a fluid – e.g., its ability to retain and transfer heat – implies the existence of microstructure. Today we know of several examples of spacetimes that exhibit thermal phenomena, which raises the following questions: Could it be that spacetime itself has a microstructure and classical gravity is just the thermodynamic limit of the statistical mechanics of these *atoms of spacetime*? If so, does *classical* gravity show evidence of this feature? Several recent results suggest that this could indeed be the case. This article describes the context, concrete results, and implications of this approach which views gravity as an emergent phenomenon.

12.1 Introduction, Motivation, and Summary

Recent results suggest that gravity could be an emergent phenomenon, like fluid mechanics or elasticity, with the field equations governing gravitational dynamics having the same status as, say, the equations of fluid mechanics. This alternative perspective for classical gravity appears to be conceptually more satisfying than the standard perspective in certain aspects [12.1–3]. This article describes several aspects of this alternate paradigm and its consequences.

The context in which such an approach arises is the following. We will assume that there are certain pregeometric variables underlying the spacetime structure and that there exists a microscopic theory describing their dynamics. If we think of continuum spacetime as conceptually analogous to a continuum description of a fluid, the microscopic – quantum gravitational – description is analogous to statistical mechanics of the molecules of the fluid. (That is, the pregeometric variables will be the *atoms of spacetime*.) The exact theory should allow us to construct, in a coarse-grained long wavelength limit, a smooth spacetime and the effective degrees of freedom (like the metric tensor) in terms of pregeometric variables. This procedure is analogous to the definition of variables like pressure, temperature, etc., for a fluid in terms of microscopic variables. The dynamical equations of the microscopic theory will also

lead to some effective description for the emergent degrees of freedom. For example, given the microscopic picture of molecules moving randomly inside a container and colliding with the walls, one can obtain the ideal gas law $PV = Nk_{\mathrm{B}}T$ governing the macroscopic (*emergent*) variables like pressure, temperature, etc. Similarly, once we have the correct theory of quantum gravity (in terms of pregeometric variables), we should be able to obtain the field equations governing the spacetime geometry by a suitable long wavelength description.

Such an approach, unfortunately, works only when we know the underlying microscopic theory. How do we proceed when we do *not* know the underlying theory, which is the current situation? In the case of normal matter, this was indeed the situation a couple of centuries back and physicists could successfully describe the behavior of matter using thermodynamic concepts, even before we understood the molecular structure of matter. The emergent paradigm for gravity is a similar attempt to describe gravity in a thermodynamic language, given the fact that we do not know the exact, microscopic description of spacetime.

In the case of normal matter, the evidence for the existence of microscopic degrees of freedom is provided by the elementary observation that matter can be heated up. Boltzmann strongly emphasized the fact that matter can store and transfer energy in the form of heat *only because* it contains microscopic degrees of freedom. If a fluid admits continuum description all the way and has no internal degrees of freedom, it cannot store heat or have a useful notion of temperature. In other words, the existence of thermal phenomena provides strong internal evidence that continuum fluid mechanics misses some essential aspect of physics. This is clearly seen, in the case of an ideal gas, if we write

$$Nk_{\mathrm{B}} = \frac{PV}{T} . \tag{12.1}$$

The quantities on the right-hand side are well defined and meaningful in thermodynamics, but the N on the left-hand side has no physical meaning in thermodynamics. It counts the number of microscopic degrees of freedom of the gas and is irrelevant in the strictly continuum limit. With hindsight, we now know that the existence of the gas constant $R = Nk_{\mathrm{B}}$ for a mole of gas has a microscopic interpretation related to the internal degrees of freedom. This is one example in which we could have guessed (as Boltzmann did) the existence of microscopic degrees of freedom – and even counted it, as the Avogadro number for a mole of gas – from the existence of thermal phenomena in normal matter.

In a similar vein, it makes sense to ask whether the continuum description of spacetime dynamics *exhibits evidence* of a more fundamental level of microscopic description in terms of *atoms of spacetime*. It would, of course, be impossible to describe the precise statistical mechanics of such microscopic degrees of freedom knowing only the continuum description of spacetime, just as one could not understand the constituents of matter in the days of Boltzmann. However, it must be possible to provide internal evidence strongly suggesting the *existence of* such atoms of spacetime by probing the continuum theories of gravity.

The first task of the emergent paradigm that is described here is to provide such *internal evidence* by closely examining classical theories of gravity. We will discuss several peculiar features in the structure of classical gravitational theories, which have no explanation within the standard framework and have to be accepted as just algebraic accidents. It turns out that these peculiar features provide us with hints about the underlying microscopic theory. In that sense, they are conceptually similar to the equality of inertial and gravitational mass (which could have been thought of as an algebraic accident in Newtonian gravity but finds a deeper explanation when gravity is treated as spacetime geometry) or the fact that matter can be heated up (which defied a fundamental explanation until Boltzmann postulated the existence of microscopic degrees of freedom). These features strongly suggest interpreting classical gravity as an emergent phenomenon with its field equations having the same status as equations of fluid mechanics or elasticity. Careful analysis of classical gravity (with one single quantum mechanical input, viz., the Davies–Unruh temperature [12.4, 5] of local Rindler horizon) leads us to this conclusion.

In the context of normal matter, a more formal link between microscopic and macroscopic description is established by specifying certain thermodynamic potentials like entropy, free-energy, etc., as suitable functions of emergent variables like pressure, temperature, etc. The functional form of these potentials can, in principle, be derived from microscopic theory but is postulated phenomenologically from the known behavior of the macroscopic systems when we do not know the microscopic theory. Similarly, if the ideas are correct, we would expect such a thermodynamic approach to work in this case of spacetime as well. It should be possible to write down, say, an entropy or free energy density for spacetime in terms of suitable variables, the

extremum of which should lead to a consistency condition on the background spacetime – which will act as the equation of motion. Such a consistency condition arises even in normal thermodynamics, though it is not often stated as such. For a gas of N molecules in a volume V, we can express the kinetic energy and momentum transfer through collisions to the walls of the container per unit area and time, entirely in terms of the microscopic variables of the molecules. Coarse graining these we obtain the macroscopic variables T and P. The equations of microscopic physics now demand the *consistency condition*

$$\frac{\begin{pmatrix}\text{mean momentum transfer to walls}\\ \text{per unit time per unit area}\end{pmatrix}}{(\text{mean kinetic energy})} = \frac{N}{V} = \frac{P}{T}\,, \tag{12.2}$$

between the two coarse-grained variables. It will be shown in Sect. 12.5 that such an approach is indeed possible to obtain the field equations of gravity.

In the case of normal matter, the laws of thermodynamics hold for all kinds of matter and do not depend on the kind of matter (ideal gas, liquid crystal, metal, ...) one is studying. The information about the *specific* kind of matter that one is studying is provided by the *specific* functional form of the thermodynamic potential say, free energy $F = F(T, V)$. Similarly, the thermodynamic framework is capable of describing a wide class of possible gravitational field equations for the effective degrees of freedom. Which of these field equations actually describe nature depends on the specific functional form of the thermodynamic potential, say, the entropy density of spacetime. It turns out that this information is encoded in a tensor P^{ab}_{cd}, which I call the *entropy tensor*, and the nature of the resulting theory depends on the dimension of spacetime. In particular, if $D = 4$, the thermodynamic paradigm selects Einstein's theory uniquely under some very reasonable assumptions.

This approach can be thought of as a *top-down* view (in real space, like zooming into a Google map of terrain!) from classical gravity to quantum gravity. Specific quantum gravitational models which approach the problem *bottom-up* have to maintain consistency with the features of classical gravity described in the sequel. In particular, this approach makes precise the task of the microscopic quantum gravity model. It should lead to a specific functional form for the entropy or free energy density of spacetime, just as microscopic statistical mechanics will lead to a specific entropy (or free energy) functional for a material system. In this sense, the emergent approach complements microscopic approaches based on toy models of quantum gravity. Such a perspective also implies that quantizing any classical gravitational field will be similar to quantizing equations of fluid dynamics or elasticity. Gravitons will be just like phonons in a solid. Neither will give us insights into the deeper microstructure (spacetime or atoms).

The alternative paradigm has important implications for cosmology and, in fact, suggests linking cosmology with the emergence of space itself in a special manner. The last part of the article will describe the implications of this paradigm for cosmology and new features that are arising [12.3].

The idea that gravity is an emergent phenomenon has a long history. *Sakharov* [12.6] was probably the first to suggest (in 1968) an analogy between spacetime dynamics and elasticity, although the way he implemented the ideas was rather restrictive. *Bei-lok Hu*, *Volovik*, and others have emphasized such points of view since the mid-1990s [12.7, 8], and a concrete procedure to obtaining Einstein's field equations from thermodynamic argument was attempted in 1995 by *Jacobson* [12.9]. These ideas have been generalized significantly to cover a much wider class of gravitational theories and have also been explored from widely different perspectives by this author and his collaborators [12.1–3] in the last decade. This chapter will concentrate on the work of the author and his collaborators.

The chapter is set out as follows: in the next section, the conventional approach to classical gravitational theories is rapidly reviewed, emphasizing some of its curious features. Section 12.3 introduces quantum mechanics into the picture through Davies–Unruh temperature and describes the notions of temperature and entropy which can be associated with null surfaces that act as horizons for a particular class of observers. In Sect. 12.4, the intimate relationship between field equations of gravity and thermodynamics of horizon is discussed from different points of view. Building on this background, Sect. 12.5 provides an alternative perspective in which the gravitational dynamics is obtained from a thermodynamic extremum principle. Last two sections describe some key new implications of the emergent paradigm for cosmology. Section 12.6 discusses how one could think of space itself as emergent in the context of cosmology and Sect. 12.7 shows how these ideas might solve one of the most puzzling problems of theoretical physics, viz. the numerical value of the cosmological constant. The final section briefly summarizes the conclusions.

12.2 Curious Features in the Conventional Approach to Classical Gravity

Given the success and elegance of general theory of relativity, it is probably important to begin by answering a question like *Why fix it when it works?* for classical gravity. So let us start by critically reviewing the conventional approach and discussing several of its peculiarities. It is useful for this purpose to distinguish the kinematical structure of gravitational theory from the dynamics which is encoded in the field equation.

12.2.1 Kinematics of Gravity and the Ubiquity of Horizons

Using a fairly natural interpretation of the principle of equivalence and the principle of general covariance, it is possible to conclude that the *kinematics* of gravity is closely linked to the spacetime structure and can be described by the metric tensor $g_{ij}(x^a)$. Given a metric, and the associated spacetime geometry, one can write down the covariant equations of motion for matter fields and work out *how gravity makes matter move*. The usual beauty and elegance attributed to Einstein's general relativity, arise from this natural *kinematic* description of gravity in terms of the geometry of spacetime. The alternative perspective that is described later retains this kinematic structure and hence *loses none of this elegance*. (In addition, as we shall see, it will describe the *dynamics* of gravity as well from a nice principle.)

Even at the level of kinematics, the geometrical description introduces two new features which have no analog in other areas of physics.

First, the principle of equivalence – along with a judicious set of thought experiments – implies that gravity influences the propagation of light and hence affects the *causal structure* of spacetime. It is possible to write down metrics $g_{ij}(x^a)$ such that there are regions in spacetime which cannot communicate with the rest of the spacetime because of the nontrivial causal structure. Unless we introduce some principle to exclude such metrics – and no such principle is likely to exist, for reasons described below – it is obvious that the amount of information accessible to different observers will be different. This does not happen in any other physical theory; in the absence of gravity one can introduce global inertial frame in flat spacetime which has a standard causal structure.

Second, the principle of general covariance implies that all observers, moving along any (nonspace-like) world line, have an equal right to study and describe physics. In flat spacetime, since there exists a global inertial frame with the metric $g_{ab} = \eta_{ab}$, it makes sense to give special status to inertial observers. Noninertial observers may see certain phenomenon which inertial observers do not see but we do have a right to treat inertial observers as special. Mathematically one can attribute *all* the difference between the actual metric g_{ab} and the flat metric η_{ab} to the choice of coordinates. However, in a curved spacetime (i. e., in the presence of gravity), there is no global inertial frame; we can no longer say *how much* of g_{ab} is due to coordinate choice and *how much* of it is due to genuine curvature. Locally, the freely falling frame (FFF) takes away the effects of the coordinate system and leaves only the imprint of curvature; but one cannot do this globally so we should be prepared to treat all observers (and their coordinate systems) as equal. Again, this does not happen in other theories; while one can use noninertial coordinates for technical convenience the global inertial frame remains special.

Combining these two features leads to an important consequence viz., *horizons are ubiquitous*. One can construct a noninertial coordinate system in flat spacetime in which a class of observers (say, for example, uniformly accelerated observers which we will call Rindler observers) will perceive a horizon and will use a nontrivial metric. These observers will view physical phenomena differently from inertial observers in flat spacetime – which should be accepted as an inevitable consequence of general covariance and the principle of equivalence. All one can wish for is a clear dictionary translating the physical phenomena as viewed by observers in different state of motion, in spite of limitations of causality implied by a nontrivial metric.

These features do not create any serious issues in *classical* physics and it is fairly easy to discuss classical phenomena in a general covariant manner. However, it turns out that quantum field theory introduces of certain amount of conceptual tension vis-a-vis the principle of general covariance. This will be one of the central themes running through this article.

12.2.2 The Troubles with Gravitational Dynamics

To complete the picture, we need some prescription for determining the form of the metric tensor at all events in spacetime. The conventional view has been to think of the metric tensor as akin to a field, write down an action principle, and obtain a differential equation that

determines the metric tensor. While such an approach proved very successful in other areas of physics, it is not at all clear – a priori – why the dynamics of the *spacetime geometry* should be derivable from an action principle.

This question brings to sharp focus the dual nature of gravity which behaves as a field as well as playing a role in determining the spacetime structure. The kinematics of gravity mostly makes use of the geometrical nature of gravity with the metric tensor being predefined on a differential manifold, etc. But an attempt to write down a variational principle to arrive at the equations which determine the evolution of the metric tensor, arises from the view that gravity also behaves *like a field*. It is entirely conceivable that the description of *spacetime* may need a totally different approach! All that one needs is *some* physical principle which leads to the necessary differential equations and, in the later sections, we will describe a viable alternative to the standard interpretation. However, for the moment, let us assume that we are interested in writing down a scalar Lagrangian that will lead to the differential equations governing the evolution of the metric.

We immediately face the difficulty that we have *no elegant governing principle* to choose such a Lagrangian. For example, if we take the view that *gravity is like a field* seriously, we will look for a Lagrangian which is quadratic in the derivatives of the metric. However, there are no scalars that can be built from the metric and its first derivatives, which is quadratic in the first derivatives, unlike in other field theories. (There are, of course, ways of getting around these problems by technical artifacts, but the fact remains that the simplest and most natural ideas run into trouble.) So, in contrast to the kinematics of gravity, the *dynamics* of gravity is crying out for a fundamental physical principle for its determination.

This should give us a warning that it may be wrong to think of gravity as a field; but let us ignore this and carry on forward. Then the simplest choice – which turns out to be adequate and even unique in a sense described below – would be to choose a Lagrangian $L\left(R^{ab}_{cd}, g^{ij}\right)$ which depends on the curvature R^{ab}_{cd} and the metric but not on the derivatives of the curvature. (Most of the conceptual comments made here will go through even if the Lagrangian depends on the derivatives of the curvature tensor.)

The next problem we face is that actions defined using such scalars do not possess a functional derivative with respect to metric; that is, we cannot have a well-defined variational principle when we fix the metric alone on the boundary of a region. Once again, we need to do something special for gravity – either impose somewhat unusual boundary conditions or add some surface term to cancel unwanted terms in the variation. If we do this, we obtain the following field equation:

$$\begin{aligned} \mathcal{G}^b_a &= P^{de}_{ac}R^{bc}_{de} - 2\nabla^c\nabla_d P^{db}_{ac} - \frac{1}{2}L\delta^b_a \\ &\equiv \mathcal{R}^b_a - \frac{1}{2}L\delta^b_a = \frac{1}{2}T^b_a\,, \\ P^{ab}_{cd} &\equiv \left(\frac{\partial L}{\partial R^{cd}_{ab}}\right), \end{aligned} \tag{12.3}$$

where T_{ab} is the stress tensor of matter. The term $\mathcal{R}_{ab}$ is actually symmetric but it is nontrivial to prove this result [12.10]. We thus see that the dynamics is encoded in the tensor P^{ab}_{cd}, which also has the symmetries of the curvature tensor. Given a particular spacetime with a certain curvature tensor, we determine its dynamics using P^{ab}_{cd} with different P^{ab}_{cd}s leading to different dynamics.

In general, (12.3) will contain fourth-order derivatives of the metric tensor and it is not clear whether one would like to allow this. In the conventional approach, when we think of the metric as akin to a field, it seems reasonable to limit ourselves to equations of motion that are second order in derivatives which can be achieved by choosing L such that $\nabla_a P^{abcd} = 0$. Interestingly enough, one can determine [12.11, 12] the *most general* scalar functionals $L(R^{ab}_{cd}, g^{ij})$ satisfying this condition. These scalars are, in fact, independent of the metric (if we express the Lagrangian as a function of R^{ab}_{cd} and g^{ab}) and can be expressed as polynomials in curvature tensor R^{ab}_{cd} contracted with a string of Kronecker delta functions in the form of determinant tensors. With this choice, we are led to the (so-called) Lanczos–Lovelock models with the field equations

$$\begin{aligned} P^{de}_{ac}R^{bc}_{de} - \frac{1}{2}L\delta^b_a &= \mathcal{R}^b_a - \frac{1}{2m}\mathcal{R}\delta^b_a = \frac{1}{2}T^b_a\,; \\ \mathcal{R}^b_a &\equiv P^{de}_{ac}R^{bc}_{de}\,, \quad \mathcal{R} = \mathcal{R}^a_a\,. \end{aligned} \tag{12.4}$$

The second form of the equation is valid for the m-th order Lanczos–Lovelock model for which $\mathcal{R} = R^{ab}_{cd}\left(\partial L/\partial R^{ab}_{cd}\right) = mL$. In the simplest context of $m = 1$ we take $L \propto R = R/16\pi$ (with conventional normalization), leading to $P^{ab}_{cd} = (32\pi)^{-1}\left(\delta^a_c\delta^b_d - \delta^a_d\delta^b_c\right)$, and $\mathcal{R}^a_b = R^a_b/16\pi$, $\mathcal{G}^a_b = G^a_b/16\pi$ so that one recovers Einstein's equations. (It is easy to see that in $D = 4$ we recover Einstein's theory *uniquely*. Thus, if one insists

that $D = 4$ and that the Lagrangian should be built from R^{ab}_{cd} and Kronecker deltas, we obtain Einstein's theory.)

This action functional for the Lanczos–Lovelock model has several peculiar features which again should warn us that maybe we have not really understood the nature of gravitational dynamics.

First, as we have already mentioned, the functional derivative of L with respect to g_{ab} does not exist, due to the presence of second derivatives of the metric. This is usually tackled by adding some surface terms. These surface terms are neither unique – a fact not usually appreciated by many, who think the York–Gibbons–Hawking surface term [12.13, 14] proportional to K is unique in GR, which it is not [12.15, 16] – nor simple for Lanczos–Lovelock models (see, e.g., [12.17]); the mere fact that we have to do it, is a strange feature of gravitational theories.

A second, and related, peculiarity is that one can separate the Lanczos–Lovelock Lagrangian into bulk and surface terms ($L = L_{\text{bulk}} + L_{\text{sur}}$) connected by a peculiar relation

$$\sqrt{-g}L_{\text{sur}} = -\partial_a \left(g_{ij} \frac{\delta\sqrt{-g}L_{\text{bulk}}}{\delta(\partial_a g_{ij})} \right), \tag{12.5}$$

thereby duplicating the information in bulk and boundary terms [12.18]. All Lanczos–Lovelock action functionals have this structure [12.19] and nobody knows why. In fact, in a small region around any event $\mathcal{P}$, the Einstein–Hilbert action reduces to a pure surface term when evaluated in the Riemann normal coordinates, suggesting that the dynamical content is actually stored on the boundary rather than in the bulk. We will keep coming across such correspondence between bulk and boundary behavior (all of which we will call *holographic*) as we go along. No such issues (bulk and boundary terms, nonexistence of functional derivative without extra prescriptions, etc.) arise in any other field theory, including nonabelian gauge theories.

Third, the fact that $R = L_{\text{bulk}} + L_{\text{sur}}$ in Einstein gravity, with the two terms being related by (12.5), suggests that Einstein–Hilbert action should be thought of as a momentum-space action [12.15, p. 292]. This is clear if we use $f^{ab} \equiv \sqrt{-g}g^{ab}$ as the dynamical variables and with the associated momenta

$$N^i_{jk} \equiv \frac{\partial(\sqrt{-g}L_{\text{bulk}})}{\partial\left(\partial_i f^{jk}\right)} = -\left[\Gamma^i_{jk} - \frac{1}{2}\left(\delta^i_j \Gamma^a_{ka} + \delta^i_k \Gamma^a_{ja} \right) \right]. \tag{12.6}$$

Then it is easy to show that

$$\delta(\sqrt{-g}R) = \sqrt{-g}G_{ab}\delta g^{ab} - \partial_i \left(f^{jk}\delta N^i_{jk} \right), \tag{12.7}$$

so that equations of motion will arise from $\delta A_{EH} = 0$ if we fix the momenta N^i_{jk} on the boundary. If we decide *not* to add any surface term to Einstein–Hilbert action, then we can still obtain the field equations if we demand

$$\delta A_{EH} = -\int_{\partial\mathcal{V}} \mathrm{d}^3x\sqrt{h}n_i g^{jk}\delta N^i_{jk}, \tag{12.8}$$

instead of the usual $\delta A_{EH} = 0$. This looks more like the change in the bulk property being equated to a change in the surface property rather than standard action principle. As we shall see later, all these features have thermodynamic interpretation.

Finally, there is another curious aspect related to the surface term in Einstein–Hilbert action. When we introduce an action functional to describe the dynamics of gravity, we are clearly relying on the idea that gravity is similar to other fields in nature and when quantized it will lead to the concept of gravitons. In standard quantum field theory action is dimensionless and all fields will have the dimension of inverse length, in natural units. In the case of a gravitational field, we associate a second rank symmetric tensor field, H_{ab}, to describe the graviton and write the metric g_{ab} as $g_{ab} = \eta_{ab} + \lambda H_{ab}$, where λ is a constant with dimensions of length. (In normal units, $\lambda^2 = 16\pi(G\hbar/c^3)$.) We can now use this expansion in Einstein–Hilbert action and retain terms up to the lowest nonvanishing order in the bulk and surface terms to obtain the action functional in the form: $\mathcal{A} \equiv \mathcal{A}_{\text{quad}} + \mathcal{A}_{\text{sur}}$. We then find that $\mathcal{A}$ matches exactly with the action for the spin-2 field known as Fierz–Pauli action (see, e.g., [12.20]) but the surface term – which is usually ignored in standard field theory – is *nonanalytic* in the coupling constant

$$\mathcal{A}_{\text{sur}} = \frac{1}{4\lambda} \int \mathrm{d}^4x \partial_a \partial_b \left[H^{ab} - \eta^{ab} H^i_i \right] + \mathcal{O}(1). \tag{12.9}$$

In fact, the nonanalytic behavior of $\mathcal{A}_{\text{sur}}$ on λ can be obtained from fairly simple considerations related to the algebraic structure of the curvature scalar. In terms of a spin-2 field, the final metric is $g_{ab} = \eta_{ab} + \lambda H_{ab}$, where $\lambda \propto \sqrt{G}$ has the dimension of length and H_{ab} has the correct dimension of $(\text{length})^{-1}$ in natural units with $\hbar = c = 1$. Since the scalar curvature has the structure

$R \simeq (\partial g)^2 + \partial^2 g$, substitution of $g_{ab} = \eta_{ab} + \lambda H_{ab}$ gives to the lowest order

$$L_{EH} \propto \frac{1}{\lambda^2} R \simeq (\partial H)^2 + \frac{1}{\lambda} \partial^2 H . \qquad (12.10)$$

Thus even the full Einstein–Hilbert Lagrangian is non-analytic in λ because of the surface term, which has no interpretation in terms of gravitons.

If we choose to ignore these peculiarities and decide to treat gravity naively as some kind of a field, then, as far as *classical* description goes, the story ends here. We may postulate $D = 4$ and work out the consequences of the theory and determine any parameters (e.g., Newton's constant and the cosmological constant) by comparing theory with observation – which is what we were taught to do in the grad school. The most serious inconsistency we will then face is that the theory is incapable of answering well-posed questions as regards some of its solutions. Like, for example, what is the fate of matter in the context of gravitational collapse to a singularity as viewed by an observer freely falling into the singularity or what happened to our universe at sufficiently early times, etc. The existence of mathematical singularities leads to a lack of predictability in the theory, showing that the theory – at the least – is incomplete. This, coupled to the fact that sources of gravity are known to obey *quantum* laws, suggest that the more complete theory could be quantum mechanical in nature.

Since all attempts to construct a quantum theory of gravity using the conventional tools of high energy physicists – which were so successful in other contexts – have failed, it makes sense to study areas of contact and conflict between gravity and quantum theory with the hope that we will obtain some clues. As we will see, such a study reemphasizes the view that one should *not* approach the dynamics of gravity as the dynamics of some kind of a field.

12.3 Quantum Theory and Spacetime Horizons

The author believes the single most important guiding principle we can use, in understanding the quantum structure of spacetime, is the *thermodynamic properties of null surfaces*. In fact, these phenomena could be considered as important as the equality of inertial and gravitational masses (which was used by Einstein to come up with the geometric description of gravity) or the fact that normal matter can store heat (which was used by Boltzmann to work out the existence of microscopic degrees of freedom in matter). Let us elaborate on this point of view.

12.3.1 Observer-Dependent Temperature of Null Surfaces

The original idea, due to Bekenstein, that black hole horizons should be attributed an entropy found strong support from the discovery of the temperature of the black hole horizon by *Hawking* [12.21, 22]. One might have thought that these are just couple of more esoteric features special to black holes except for the discovery by *Davies* and *Unruh* [12.4, 5] (and the work of many others later) which showed that even Rindler observers in flat spacetime will attribute temperatures to the horizons they perceive. In fact, the situation is more general because one could introduce the notion of *local Rindler observers* around any event in any spacetime along the following lines.

Take any event $\mathcal{P}$ in any spacetime and construct the Riemann normal coordinates (X^i) around that event as the origin so that $g_{ab} = \eta_{ab} + \mathcal{O}(X^2)$. Observers at X = constant are locally inertial observers around $\mathcal{P}$. We can now construct local Rindler observers (and the corresponding local Rindler frame, LRF, with coordinates x^i) who move, say, with an acceleration κ along the X direction. These observers will perceive the null surface $X = T$ as a local Rindler horizon and will attribute to it a temperature $\kappa/2\pi$.

The existence of such a *local* description can be easily understood by analytically continuing the metric around $\mathcal{P}$ into the Euclidean sector. The null surfaces $X^2 - T^2 = 0$ will map to the origin of the Euclidean $T_E - X$ plane and the Rindler observers (following $x =$ constant world lines) will have Euclidean trajectories $X^2 + T_E^2 =$ constant, which are circles around the origin. The Euclidean Rindler time coordinate t_E will be periodic with a period $(2\pi/\kappa)$. Thermal phenomena of approximately local nature will arise as long as the acceleration does not change significantly over this period of the Euclidean time; this translates to the condition $\dot{\kappa}/\kappa^2 \ll 1$, which *can always be achieved by choosing sufficiently large* κ. Thus, observers close to the Euclidean origin, orbiting on circles of very small radius, will provide a local description of the thermal phenomena. (The Euclidean description of null surfaces has another advantage. Since the region beyond the hori-

zon is not accessible to the local Rindler observer, it seems appropriate to construct an effective field theory for this observer in a spacetime which has only the region accessible to it. The inaccessible region behind the null surfaces collapses to a point at the origin in the Euclidean description leaving only the region accessible to the local Rindler observer for the study of physical phenomena.) The nature of the geometry far away from $\mathcal{P}$ becomes irrelevant in the limit of $\kappa \to \infty$.

The same conclusions can also be reached by analyzing an observer close to its event horizon of, say, a Schwarzschild spacetime. With a suitable coordinate choice, the Schwarzschild metric can be approximated as a Rindler metric near the horizon, with κ replaced by the surface gravity of the black hole. An observer very close to the event horizon, performing quasi-local experiments (at length scales that are small compared to curvature scale) has no way of distinguishing between a Rindler coordinate system in a flat spacetime and the black hole spacetime, because the results of quasi-local observations performed by an observer should not depend on the nature of the geometry far away. It follows that local Rindler observers *must* attribute to their horizons the standard thermodynamic properties if black hole horizons exhibit thermal properties. This argument also shows that the *local Rindler observers will attribute an entropy density to the Rindler horizon* – which is just a null surface in flat spacetime – if black hole horizons are attributed an entropy density. What matters are the operationally well-defined, quasilocal observations, by which one cannot distinguish the thermodynamic features of Rindler horizon in flat spacetime from the event horizon of black holes. Freely falling observers will see nothing special while crossing *either* horizon, while observers accelerated with respect to the FFF will attribute thermal properties to *both* horizons.

Thus combining the principles of quantum theory (in the form of the Davies–Unruh effect in local Rindler horizons) with standard description of gravity leads to associating an *observer-dependent temperature, entropy density, etc., to all null surfaces* in spacetime. So spacetimes, like matter, appear to be hot to some observers. Let us explore the consequences of this.

12.3.2 Observer-Dependent Entropy of Null Surfaces

Given a particular metric which has a horizon with respect to a certain class of observers, one can work out the quantum field theory in that spacetime and determine the temperature of the horizon. For slowly varying horizons (with $\dot{\kappa}/\kappa^2 \ll 1$) such an analysis will lead to a temperature $\kappa/2\pi$. This result has nothing to do with the *dynamics* of gravity and it does *not* care about the field equations (if any) for which the given metric arises as a solution. In fact, once we approximate a nonextremal, slowly varying horizon as a Rindler horizon, the results translate to those which we know in flat spacetime itself and thus cannot depend on the field equations. This is to be expected because, even in the case of normal matter, the temperature contains very little information about the structure of the matter heated to that temperature.

One might have thought that the analysis that leads to temperature will also lead to an expression for entropy which is independent of the theory. Indeed, there *exists* an entropy $S = -\rho \log \rho$ (called the entanglement entropy) associated with the thermal density matrix $\rho \propto \exp(-\beta H)$ of the *matter* field in the presence of the horizon. It turns out, however, that this is *not* the entropy associated with the horizon, for two reasons. To begin with it is divergent and hence its value depends on the cut-off used; so it is useless for predicting anything. Second, the entanglement entropy is always proportional to the area of the horizon but the correct entropy (which will obey the appropriate laws of black hole physics, for example) is *not* proportional to the area except in Einstein's theory. (The situation is slightly different in the emergent paradigm where one can argue that the regularization procedure needs to be modified *but in a Lorentz invariant manner*. Then, using a generalization of ideas described in [12.23–27], one can possibly tackle this issue. We will not this discuss here; for more details, see [12.28]) *The correct entropy of a horizon depends on the theory and arises in a manner which defies simple interpretation in the conventional approach.* There are two mathematically well-defined procedures for computing the correct entropy of horizons and I will now describe them. In the conventional approach, we have no idea why either procedure should lead to a *thermodynamic* quantity.

Entropy from Diffeomorphism Invariance

In the first method, one proceeds in the following manner [12.29, 30]. In any theory with a generally covariant action, the invariance of the action under infinitesimal coordinate transformation $x^a \to x^a + q^a(x)$ leads to the conservation of a Noether current J^a related to the Noether potential J^{ab} (which depends on q^a) by $J^a \equiv \nabla_b J^{ab}$. In the case of the Lanczos–Lovelock mod-

els, these are given by

$$J^{ab} = 2P^{abcd}\nabla_c q_d\,,$$
$$J^a = 2P^{abcd}\nabla_b\nabla_c q_d\,. \tag{12.11}$$

The entropy of the horizon is then given by the surface integral

$$S_{\text{Noether}} \equiv \frac{1}{T}\int \mathrm{d}^{D-2}\Sigma_{ab}\,J^{ab} = \frac{1}{4}\oint_{\mathcal{H}} \left(32\pi P^{ab}_{cd}\right)\epsilon_{ab}\epsilon^{dc}\,\mathrm{d}\sigma\,, \tag{12.12}$$

where $T = \beta^{-1} = \kappa/2\pi$ is the horizon temperature and $q^a = \xi^a$, where ξ^a is the local Killing vector corresponding to time translation symmetry of the LRF. In the final expression the integral is over any surface with $(D-2)$ dimension, which is a space-like cross-section of the Killing horizon on which the norm of ξ^a vanishes, with ϵ_{ab} denoting the bivector normal to the bifurcation surface.

In Einstein's theory, with $32\pi P^{ab}_{cd} = \left(\delta^a_c\delta^b_d - \delta^a_d\delta^b_c\right)$, the entropy will be one quarter of the area of the horizon. However, in general, the entropy of the horizon is *not* proportional to the area and depends on the theory. (Even in Einstein's theory, the thermodynamical variables T and S have strange limiting behavior which is not well understood. For example, the flat spacetime can be thought of as the $M \to 0$ of the Schwarzschild metric or as the $H \to 0$ limit of de Sitter spacetime. In the first case, the entropy $S = 4\pi M^2$ vanishes as to be expected for flat spacetime but the temperature $T = (1/8\pi M)$ diverges. In the second case, the temperature $T = (H/2\pi)$ does vanish but the entropy $S = \pi/H^2$ diverges. These features probably indicate the nonperturbative nature of spacetimes with horizons when considered as excitations of the gravitational vacuum represented by flat spacetime). This feature again shows that, as mentioned before, the entanglement entropy cannot be identified with the entropy of the Lanczos–Lovelock models, since the horizon entropy is given in terms of P^{ab}_{cd}, which we may call the *entropy tensor* of the theory.

The knowledge of the functional dependence of S on ϵ_{ab} (or the dependence of J^{ab} on $\nabla_i q_j$), say, is equivalent to the knowledge of P^{ab}_{cd} and – consequently – the field equations of the theory through (12.4). One could think of spacetime having two tensors R^{ab}_{cd} and P^{ab}_{cd} associated with it. The first one describes curvature while the second one describes the entropy of null surfaces. These two tensors are related by $P^{ab}_{cd} = \partial L/\partial R^{cd}_{ab}$ which is reminiscent of thermodynamic duals with L being some thermodynamic potential. The field equations, (12.4), of the theory are determined by the product of entropy tensor and curvature tensor $\mathcal{R}^b_a \equiv P^{de}_{ac}R^{bc}_{de}$ so that different entropy tensors P^{ab}_{cd} will lead to different field equations for the same spacetime geometry. This seems to give a nice separation of the dynamics of spacetime and encode it in its entropy.

All these features – in particular why *diffeomorphism invariance* should have anything to do with a *thermodynamic quantity* like the horizon entropy – are mysterious in conventional approaches but we will see later that all these ideas fit naturally with the emergent perspective.

Entropy from the Surface Term in the Action Functional

There is an alternative way of computing the same horizon entropy – from the surface term of the gravitational action – which also defies physical interpretation in the conventional approach. Recall that the field equations can be obtained by varying only the bulk term (e.g., Γ^2 term in Einstein's theory) in the action ignoring (or by canceling with a counterterm) the surface term in the action. However, if we evaluate the surface term on the horizon of any solution to the field equations of the theory, we obtains the entropy of the horizon when we fix the range of time integration using the periodicity in the Euclidean time!

For example, in Einstein's theory, we have

$$16\pi L_{\text{sur}} = \partial_c(\sqrt{-g}V^c)$$

with [12.15, eq (6.15)]

$$V^c = -(1/g)\partial_b\left(gg^{bc}\right),$$

while the Gibbons–Hawking–York counterterm is the integral of $K/8\pi$ over the surface. If we use a Rindler approximation to the near horizon metric (with $-g_{00} = 1/g_{xx} = N^2 = 2\kappa x$ and evaluate these on $N =$ const surface we will obtain

$$\frac{1}{8\pi}\int_x \mathrm{d}t\,\mathrm{d}^2x_\perp\sqrt{h}K = \frac{1}{16\pi}\int_x \mathrm{d}t\,\mathrm{d}^2x_\perp V^x = \pm t\left(\frac{\kappa A_\perp}{8\pi}\right), \tag{12.13}$$

where $A_\perp$ is the transverse area. (The sign depends on the convention chosen for the outward normal or

whether the contribution of the integral is taken at the inner or outer boundaries; see e.g., the discussion in [12.31, 32].) In the Euclidean sector the range of time integration is $(0, 2\pi/\kappa)$, which leads to, with a proper choice of sign,

$$\mathcal{A}^E_{\text{sur}} = \frac{1}{4} A_\perp\,, \tag{12.14}$$

which is the entropy. More generally, a static, near-horizon geometry can be described by the metric [12.33–35]

$$\mathrm{d}s^2 = -N^2\,\mathrm{d}t^2 + \mathrm{d}l^2 + \sigma_{AB}\,\mathrm{d}x^A\,\mathrm{d}x^B\,, \tag{12.15}$$

where N and σ_{AB} have near-horizon behavior of the form

$$N = \kappa l + O(l^3)\,; \quad \sigma_{AB} = \mu_{AB}(x^A) + O(l^2)\,, \tag{12.16}$$

where $l = 0$ is taken to be the location of the horizon. The integrals in (12.13) again lead to the same result.

This raises the question:

How does the surface term, which was discarded before the field equations were even obtained, know about the entropy associated with a solution to those field equations?!

The only explanation seems to lie in the duplication of information between surface and bulk terms described by the relation in (12.5). However, if a part of the action functional is the entropy, it makes sense to look for a thermodynamic interpretation to the full action functional! So maybe we have been deriving field equations by extremizing a thermodynamic potential rather than action – a point of view that we will come back to.

Incidentally, note that (12.13) allows us to define a *surface Hamiltonian* for a horizon [12.36]. In the Rindler limit the integrand does not depend on t, y, z, and hence the result of integration must be proportional to $tA_\perp$, and we only need to determine the numerical factor of proportionality. Choosing the minus sign in (12.13), we can define the horizon surface Hamiltonian as

$$\begin{aligned} H_{\text{sur}} &\equiv -\frac{\partial \mathcal{A}_{\text{sur}}}{\partial t} = \frac{1}{8\pi}\int_x \mathrm{d}^2x_\perp \sqrt{h}K \\ &= \left(\frac{\kappa A_\perp}{8\pi}\right) = TS\,. \end{aligned} \tag{12.17}$$

This Hamiltonian plays an interesting role in the study of black hole horizons [12.31, 32] and is closely related to the phase of the semiclassical wave function of the black hole. When a semiclassical black hole is in contact with external matter fields, the probability for its area to change by $\Delta A_\perp$ is governed by a Fourier transform of the form

$$\begin{aligned} \mathcal{P}(\Delta A_\perp) &= \int_{-\infty}^{\infty} \mathrm{d}t F_m(t)\exp[-\mathrm{i}t\Delta H_{\text{sur}}] \\ &= \int_{-\infty}^{\infty} \mathrm{d}t F_m(t)\exp\left[-\mathrm{i}t\frac{\kappa}{8\pi}\Delta A_\perp\right]\,, \end{aligned} \tag{12.18}$$

where $F_m(t)$ is a suitable matter variable. Because of the exponential redshift near the horizon, the time evolution of $F_m(t)$ will have the asymptotic form $\exp[-\mathrm{i}C\exp(-\kappa t)]$ with some constant C. This will lead to the result that the relative probability for black hole radiation changing its area by $\Delta A_\perp$ is given by $\exp[\Delta A_\perp/4]$.

One can think of H_{sur} as the heat content of the horizon in the emergent perspective because it satisfies the relation $\mathrm{d}S = \mathrm{d}H_{\text{sur}}/T$. The corresponding horizon heat energy per unit area of the horizon, $H_{\text{sur}}/A_\perp = \kappa/8\pi = P$ appears as the pressure term in the Navier–Stokes equation obtained by projecting Einstein's equation onto the null surface [12.37, 38] and leads to the equation of state $PA = TS$ (Sect. 12.4.1). This heat energy per unit area of the horizon, taken to be $x^1 = \text{const}$ surface with $n_c = \delta^1_c$, is

$$\begin{aligned} \mathcal{H} &= \frac{NK}{8\pi} = \frac{1}{16\pi}\sqrt{-g}V^c n_c \\ &= -\frac{1}{16\pi}\sqrt{-g}n_c\left(g^{ab}N^c_{ab}\right) \end{aligned} \tag{12.19}$$

(with a suitable choice of signs), showing that it is also closely related to gravitational momentum density defined in (12.6).

The existence of horizon entropy is a nonperturbative result [12.20] and has no interpretation in terms of gravitons. We saw earlier that the surface term is nonanalytic in the coupling constant when we write the metric in terms of a spin-2 graviton field as $g_{ab} = \eta_{ab} + \lambda H_{ab}$ with $\lambda^2 = 16\pi(G\hbar/c^3)$. Therefore, we cannot interpret the surface term – and hence – the horizon entropy (which, as we have seen, can be obtained from the surface term in the action) in the linear, weak coupling limit of gravity. The fact that horizon degrees of freedom which contribute to the entropy are not connected

with gravitons is also obvious from another fact. There are black hole solutions in 1+2-dimensional gravity with a sensible entropy and thermodynamics. However, in 1+2 dimension there are no propagating degrees of freedom or gravitons.

We thus have two different procedures for computing the entropy of the horizon. The first one described in *Entropy from Diffeomorphism Invariance* uses the Noether current related to diffeomorphism, while the approach developed in this section gives us the surface term of the action functional. Curiously enough, there exists a simple connection [12.36] between these two ways of computing the entropy, which does not seem to have been emphasized in the literature. The Gibbons–Hawking–York surface term in general relativity can *also* be written as a *volume integral*

$$\begin{aligned}\mathcal{A}_{\mathrm{sur}} &= \frac{1}{8\pi}\int_{\partial\mathcal{V}} \sqrt{h}\,\mathrm{d}^3x K \\ &= \frac{1}{8\pi}\int_{\mathcal{V}} \sqrt{-g}\,\mathrm{d}^4x \nabla_a(Kn^a)\,,\end{aligned} \tag{12.20}$$

where n^a is any vector which coincides with the unit normal to the boundary $\partial\mathcal{V}$ of the region $\mathcal{V}$ and $K = -\nabla_a n^a$. Since this expression is a scalar, it also leads to a conserved Noether current $J^a \equiv \nabla_b J^{ab}$ corresponding to the diffeomorphism $x^a \to x^a + \xi^a$. The Noether potential J^{ab} in this case (see, e.g., [12.39, Appendix]) is given by

$$J^{ab} = \frac{K}{8\pi}\left(\xi^a n^b - \xi^b n^a\right). \tag{12.21}$$

An elementary calculation in the LRF now shows that the Noether charge is given by

$$\int \mathrm{d}^{D-2}\Sigma_{ab}J^{ab} = \frac{\kappa A_\perp}{8\pi} = TS = H_{\mathrm{sur}}\,. \tag{12.22}$$

In other words, the surface Hamiltonian defined earlier in (12.17) is the same as the Noether charge for a current obtained from the surface term of the action [12.36]. It follows that the entropy corresponding to this Noether charge, given by (12.12), is the standard entropy of the horizon

$$S = \frac{1}{T}\int \mathrm{d}^{D-2}\Sigma_{ab}J^{ab} = \frac{A_\perp}{4}\,. \tag{12.23}$$

This provides a direct link between evaluation of the entropy by the boundary term in the action or from Noether current; if we use the Noether charge *corresponding to the boundary term* we obtain the correct result. As a bonus, we also see that the boundary Hamiltonian is the same as the Noether charge.

The connection between a conserved current arising from the *diffeomorphism invariance* under $x^i \to x^i + q^i$ and a *thermodynamic variable*-like entropy is yet another mystery that defies explanation in the conventional approach and is intimately related to several other peculiarities to which we have been alluding.

Gravitational Action Functional as the Free Energy of Spacetime

We mentioned earlier that, since the surface term of the action gives horizon entropy, the full gravitational Lagrangian itself is likely to have a direct thermodynamic interpretation. The Noether potential allows us to interpret it as the free energy density in any static spacetime with horizon. For any Lanczos–Lovelock model we have the result (obtained by writing the time component of the Noether current for the Killing vector $q^a = \xi^a = (1, \mathbf{0})$)

$$L = \frac{1}{\sqrt{-g}}\partial_\alpha\left(\sqrt{-g}J^{0\alpha}\right) - 2\mathcal{G}^0_0\,. \tag{12.24}$$

Only spatial derivatives contribute in the first term on the right-hand side when the spacetime is static. Integrating $L\sqrt{-g}$ over a spacetime region with time integration restricted to the interval $(0, \beta)$ to obtain the action, it is easy to see (using (12.12)) that the first term gives the entropy and the second term can be interpreted as energy [12.40]. Taking the thermodynamic interpretation as fundamental, one could even argue that all gravitational actions have a surface and bulk terms *because* they give the entropy and energy of a static spacetime with horizons, adding up to the bulk term to make the action the free energy of the spacetime. (This is closely related to the more general result in (12.5) which holds in general without the assumption of static spacetime.)

This thermodynamic interpretation of the action is reinforced by a path integral analysis. Consider the Euclidean path integral of $\exp[-A_{\mathrm{grav}}]$ over a restricted class of static, spherically symmetric geometries containing a horizon in a Lanczos–Lovelock model. This path integral can actually be performed and the resulting partition function has the form

$$Z = \sum_g \exp[-A_{\mathrm{grav}}] \propto \exp[S - \beta E]\,, \tag{12.25}$$

where S, E are the entropy and energy of the horizon and β^{-1} its temperature. This result, originally obtained in Einstein's theory [12.41], holds for all Lanczos–Lovelock models [12.42] with S and E matching with the corresponding expressions obtained by other methods.

This duplication of information in (12.5) also allows one to obtain the full action [12.43] from the surface term alone in the following manner. Let us consider the full action obtained from integrating $\sqrt{-g}(L_{\text{sur}} + L_{\text{bulk}})$ with the two terms related by (12.5). Since L_{bulk} is quadratic in the first derivative of the metric, the expression in the bracket on the right-hand side of (12.5) is linear in the first derivatives of the metric. The most general linear term of this kind can be expressed as a sum $c_1 g^{bc}\Gamma^a_{bc} + c_2 g^{ab}\Gamma^c_{bc}$. The ratio (c_2/c_1) can be fixed by demanding that this surface term should give an entropy proportional to the area of a horizon in the Rindler approximation. Integrating (12.5) and using the fact that the Rindler metric should be a solution to the field equation will then lead to [12.43] the standard expression for L_{bulk}. It is also possible to construct a specific variational principle and obtain the field equations, purely from the surface term [12.44, 45]. More importantly, since the variation of the surface term gives the change in the gravitational entropy, we can see that $\mathcal{R}^{ab}$ essentially determines the gravitational entropy density of the spacetime. We will say more about this later on.

Observer Dependence of *All* Thermodynamics

One striking conclusion we can draw from the above results is that *all* thermodynamic phenomena (including those of normal matter like a glass of water or a metal rod) must be observer-dependent. This follows immediately from the fact that the temperature attributed to the same vacuum state by an inertial observer and a Rindler observer is different; the former is zero while the latter is nonzero. If we now construct highly excited states of the vacuum (thereby making, say, a glass of water) by operating on the vacuum state with standard creation operators, the inertial and Rindler observers will attribute different temperatures to a glass of water as well. This, of course, is not of any practical relevance but assumes significance in the context of spacetime physics.

As an important aside, let us emphasize a new *principle of equivalence*, which has been brought about by these results. Consider some temperature sensitive device, say a microchip with circuits embedded in which one can measure thermal noise. If we move this microchip in different trajectories it will show a different amount of thermal noise and we can choose a trajectory in which the thermal noise is minimum. If we also check the acceleration of the microchip in these trajectories, we will find that the thermal noise is minimal when the acceleration is zero! That is, we can define the inertial motion of microchip either as one in which its acceleration is zero or the one in which it exhibits minimal thermal noise. This equivalence is highly nontrivial (and not understood at a deeper level) and arises from the mathematical similarity of vacuum fluctuations and thermal fluctuations. So we have a purely thermodynamic way of determining the geodesics of a spacetime. In a general situation we get a mix of *acceleration thermodynamics* and standard *coarse-grained thermodynamics*.

We now need to treat the *entropy of a system as an observer-dependent quantity*. A local Rindler observer will attribute an entropy density to a null surface which it perceives as a horizon, while an inertial observer will not attribute any entropy or temperature to it. The *same result holds for a black hole horizon*. A freely falling observer crossing the horizon will not attribute any special thermodynamic properties to it, while a static observer hovering outside the horizon will attribute a temperature and entropy to the horizon. We are accustomed to thinking of degrees of freedom (and resultant entropy) as an absolute quantity independent of the observer. The examples we discussed above shows that this is simply not true.

In the light of this, we next conclude that the often asked (and sometimes even answered!) question:

> *What are the degrees of freedom that contribute to the entropy of black hole horizon?*

cannot have an observer-independent answer! We need to introduce the notion of effective degrees of freedom appropriate for each observer which arises along the following lines. The full theory of gravity which we consider is invariant under a very large class of diffeomorphisms, $x^i \to x^i + q^i(x)$ for vector fields $q^i(x)$. However, when we consider a specific class of observers who perceives a null surface as a horizon, we should introduce a restricted class of diffeomorphisms which preserves the form of the metric near the null surface. Such a restriction upgrades some of the original gauge degrees of freedom (that could have been eliminated by diffeomorphisms which we are now disallowing) to effective (true) degrees of freedom as far as this particular class of observers is concerned. The entropy that these observers attribute to the null sur-

face are related to *these* degrees of freedom which may not have any relevance for, say, freely falling observers around that event. (One possible way of implementing this idea and obtaining the entropy of the horizons is explored in [12.39].)

These ideas are also important in understanding the interplay between horizon temperature and *usual* temperature of matter. Consider a box of gas at rest in an inertial coordinate system (X = constant) with the usual temperature and usual entropy which scales as the *volume* of the box. When the world line of this box crosses the null surface $X = T$, the inertial observer will see nothing peculiar. However, a Rindler observer will find that the box hovers around $X = T$ for an infinite amount of Rindler time and never crosses it! This will allow the degrees of freedom of gas to come into thermal equilibrium with the horizon degrees of freedom as far as the Rindler observer is concerned. Further, it will appear to the Rindler observer that the entropy will scale as the transverse (yz-plane) *area* of the box [12.40]. These are some of the peculiarities which arise due to observer dependence of thermodynamics.

12.4 Gravitational Dynamics and Thermodynamics of Null Surfaces

All these results suggest a close relationship between horizon thermodynamics and gravitational dynamics. Direct evidence for this relationship arises from the fact that gravitational field equations can be interpreted as thermodynamic/fluid mechanical equations. We will now discuss these results.

12.4.1 Field Equations as Thermodynamic Relations

To begin with, it can be shown that [12.35] the field equations in any Lanczos–Lovelock model, when evaluated on a static solution of the theory which has a horizon, can be expressed in the form of a thermodynamic identity $T\,\mathrm{d}S = \mathrm{d}E_g + P\,\mathrm{d}V$. Here S is the correct Wald entropy of the horizon in the theory, E_g is a *geometric expression* involving an integral of the scalar curvature of the submanifold of the horizon and $P\,\mathrm{d}V$ represents the work function of the matter source. The differentials $\mathrm{d}S$, $\mathrm{d}E_g$, etc., should be thought of as indicating the difference in S, E_g, etc., between two solutions in which the location of the horizon is infinitesimally displaced. (For a sample of related results see [12.46–54]).

To see this result in the simplest context [12.41], let us consider a static, spherically symmetric spacetime with a horizon, described by a metric

$$\mathrm{d}s^2 = -f(r)c^2\,\mathrm{d}t^2 + f^{-1}(r)\,\mathrm{d}r^2 + r^2\,\mathrm{d}\Omega^2 \,, \quad (12.26)$$

where we are using normal units temporarily. The location of the horizon is the radius $r = a$ at which the function $f(r)$ vanishes, so that $f(a) = 0$. Using the Taylor series expansion of $f(r)$ near the horizon as $f(r) \approx f'(a)(r-a)$ one can easily show that the surface gravity at the horizon is $\kappa = (c^2/2)f'(a)$. Therefore, we can associate a temperature

$$k_\mathrm{B}T = \frac{\hbar c f'(a)}{4\pi} \,, \quad (12.27)$$

with the horizon. This temperature knows nothing about the dynamics of gravity or Einstein's field equations.

Let us next write down the Einstein equation for the metric in (12.26), which is given by $(1-f) - rf'(r) = -(8\pi G/c^4)Pr^2$, where $P = T^r_r$ is the radial pressure of the matter source. When evaluated on the horizon $r = a$ this equation becomes

$$\frac{c^4}{G}\left[\frac{1}{2}f'(a)a - \frac{1}{2}\right] = 4\pi P a^2 \,. \quad (12.28)$$

This equation, which is just a textbook result, does not appear to be very thermodynamic! To see its hidden structure, consider two solutions to Einstein's equations differing infinitesimally in the parameters such that horizons occur at two different radii a and $a + \mathrm{d}a$. If we multiply (12.28) by $\mathrm{d}a$, we obtain

$$\frac{c^4}{2G}f'(a)a\,\mathrm{d}a - \frac{c^4}{2G}\mathrm{d}a = P(4\pi a^2\,\mathrm{d}a) \,. \quad (12.29)$$

The right-hand side is just $P\mathrm{d}V$ where $V = (4\pi/3)a^3$ is what is called the areal volume which is the relevant quantity to use while considering the action of pressure on a surface area. In the first term, $f'(a)$ is proportional to the horizon temperature in (12.27), and we can rewrite this term in terms of T by introducing a $\hbar$ factor (*by hand*, into an otherwise classical equation) to bring in the horizon temperature. We then find that

(12.29) reduces to

$$\underbrace{\frac{\hbar c f'(a)}{4\pi}}_{k_B T}\underbrace{\frac{c^3}{G\hbar}\mathrm{d}\left(\frac{1}{4}4\pi a^2\right)}_{\mathrm{d}S}\underbrace{-\frac{1}{2}\frac{c^4\,\mathrm{d}a}{G}}_{-\mathrm{d}E} = \underbrace{P\mathrm{d}\left(\frac{4\pi}{3}a^3\right)}_{P\mathrm{d}V} . \tag{12.30}$$

Each of the terms has a natural – and unique – thermodynamic interpretation as indicated by the labels. Thus the gravitational field equation, evaluated on the horizon now becomes the thermodynamic identity $T\mathrm{d}S = \mathrm{d}E + P\mathrm{d}V$, allowing us to read off the expressions for entropy and energy

$$S = \frac{1}{4L_P^2}(4\pi a^2) = \frac{1}{4}\frac{A_H}{L_P^2} ; \qquad E = \frac{c^4}{2G}a = \frac{c^4}{G}\left(\frac{A_H}{16\pi}\right)^{1/2} . \tag{12.31}$$

Here A_H is the horizon area and $L_P^2 = G\hbar/c^3$ is the square of the Planck length.

It is well known that black holes satisfy a set of laws similar to laws of thermodynamics, including the first law. The result derived above has a superficial similarity to it; however, the above result is *quite different* from the standard first law of black hole dynamics. For example, the usual first law of black hole mechanics will become, in the present context, $T\mathrm{d}S = \mathrm{d}\mathcal{E}$ while we have an extra term $P\mathrm{d}V$. Further, the energy $\mathcal{E}$ used in the conventional first law is defined in terms of matter source while E in our relation is purely geometrical. (For a detailed discussion of these differences, see [12.55].) The most important difference is that our result is local and does not use any property of the spacetime metric away from the horizon. Because we did not use any global notion (like asymptotical flatness), the *same result holds even for a cosmological horizon like the de Sitter horizon* once we take into the fact that we are sitting *inside* the de Sitter horizon [12.41]. In this case, we obtain the temperature and entropy of the de Sitter spacetime to be

$$k_B T = \frac{\hbar H}{2\pi} ; \quad S = \frac{\pi c^2}{L_P^2 H^2} . \tag{12.32}$$

This result also generalizes to other Friedmann universes (when H is not a constant) and gives sensible results; we will discuss these aspects in Sect. 12.6.

Classical field equations, of course, have no $\hbar$ in them, while the Davies–Unruh temperature does. However, the Davies–Unruh temperature scales as $\hbar$, while the entropy scales as $1/\hbar$ (coming from the inverse Planck area), thereby making $T\mathrm{d}S$ independent of $\hbar$! This is conceptually similar to the fact that, in normal thermodynamics, $T \propto 1/k_B, S \propto k_B$ making $T\mathrm{d}S$ independent of k_B. In both cases, the effects due to possible microstructure (indicated by nonzero $\hbar$ or k_B) disappears in the continuum limit thermodynamics.

Einstein's Equations are Navier–Stokes Equations

The discussion above dealt with *static* spacetimes which are analogous to states of a system in thermodynamic equilibrium differing in the numerical values of some parameters. What happens when we consider time-dependent situations? One can again establish a correspondence between gravity and fluid dynamics, even in the most general case. It turns out that Einstein's field equations, when projected onto *any* null surface in *any* spacetime, reduce to the form of Navier–Stokes equations in suitable variables [12.37, 38]. This result was originally known in the context of black hole spacetimes [12.56, 57] and is now generalized to any null surface perceived as a local horizon by suitable observers. Probably this is the most curious fact about the structure of the Einstein field equation, which has no explanation in conventional approaches.

Field Equations as an Entropy Balance Law on Null Surfaces

We said before that the connection between entropy and diffeomorphism invariance is a mystery in the conventional approach. However, if we interpret (from the *active* point of view) the diffeomorphism $x_i \to x^i + q^i$ as shifting (virtually) the location of null surfaces and thus the information accessible to specific observers, then the connection with entropy can be related to the cost of gravitational entropy involved in the virtual displacements of null horizons [12.58].

Consider an infinitesimal displacement of a local patch of the stretched (local Rindler) horizon $\mathcal{H}$ in the direction of its normal r_a, by an infinitesimal proper distance ϵ, which will change the proper volume by $\mathrm{d}V_{\mathrm{prop}} = \epsilon\sqrt{\sigma}\,\mathrm{d}^{D-2}x$, where σ_{ab} is the metric in the transverse space. The flux of energy through the surface will be $T^a_b\xi^b r_a$ (where ξ^a is the approximate Killing vector corresponding to translation in the local Rindler time), and the corresponding entropy flux can be obtained by multiplying the energy flux by $\beta_{\mathrm{loc}} = N\beta$.

Hence the *loss* of matter entropy to the outside observer because the virtual displacement of the horizon has engulfed some matter is

$$\delta S_m = \beta_{\rm loc}\delta E = \beta_{\rm loc} T^{aj}\xi_a r_j \, dV_{\rm prop} \,. \tag{12.33}$$

Interpreting $\beta_{\rm loc}J^a$ as the relevant gravitational entropy current, the change in the gravitational entropy is given by

$$\delta S_{\rm grav} \equiv \beta_{\rm loc} r_a J^a \, dV_{\rm prop} \,, \tag{12.34}$$

where J^a is the Noether current corresponding to the local Killing vector ξ^a given by $J^a = 2G^a_b\xi^b + L\xi^a$. (Note the appearance of the local, redshifted, temperature through $\beta_{\rm loc} = N\beta$ in both expressions.) As the stretched horizon approaches the true horizon, $Nr^a \to \xi^a$ and $\beta\xi^a\xi_a L \to 0$. Hence we obtain, in this limit, $\delta S_{\rm grav} \equiv \beta\xi_a J^a \, dV_{\rm prop} = 2\beta G^{aj}\xi_a\xi_j \, dV_{\rm prop}$. Comparing $\delta S_{\rm grav}$ and δS_m we see that the field equations $2G^a_b = T^a_b$ can be interpreted as the entropy balance condition $\delta S_{\rm grav} = \delta S_{\rm matt}$, thereby providing direct thermodynamic interpretation of the field equations as local entropy balance in LRF.

Although we work with entropy density, the factor $\beta = 2\pi/\kappa$ cancels out in this analysis – as it should, since the local Rindler observer with a specific κ was introduced only for interpretational convenience – and the relation $T\delta S_m = T\delta S_{\rm grav}$ would have served the same purpose. The expression on the right-hand side is the change in the horizon (*heat*) energy $H_{\rm sur} = TS$ of the horizon (see (12.17)) due to injection of matter energy. The context we consider corresponds to treating the local Rindler horizon as a physical system (like a hot metal plate) at a given temperature and possessing certain intrinsic degrees of freedom. Then one can integrate $\delta S = \delta E/T$ at constant T to relate change in horizon energy to injected matter energy. Any energy injected onto a null surface appears [12.1, 2] to hover just outside the horizon for a very long time as far as the local Rindler observer is concerned and thermalizes at the temperature of the horizon if it is assumed to have been held fixed. This is a local version of the well-known phenomenon that the energy dropped into a Schwarzschild black hole horizon hovers just outside $R = 2M$ as far as an outside observer is concerned. In the case of a LRF, similar effects will occur as long as the Rindler acceleration is sufficiently high; that is, if $\dot{\kappa}/\kappa^2 \ll 1$. I stress that *all these results hold for a general Lanczos–Lovelock model.*

These ideas take an interesting form in the context of cosmology. With future applications in mind, we will describe the form of entropy balance relation in the context of cosmology. Consider a Friedmann universe with expansion factor $a(t)$ and let $H(t) = \dot{a}/a$. We will assume that the surface with radius H^{-1} (in units with $c = 1, k_{\rm B} = 1$) is endowed with the entropy $S = (A/4L_{\rm P}^2) = (\pi/H^2L_{\rm P}^2)$ and temperature $T = \hbar H/2\pi$. During the time interval dt, the change of gravitational entropy is $dS/dt = (1/4L_{\rm P}^2)(dA/dt)$ and the corresponding heat flux is $T(dS/dt) = (H/8\pi G)(dA/dt)$. On the other hand, the Gibbs–Duhem relation tells us that for *matter* in the universe, the entropy density is $s_m = (1/T)(\rho + P)$ and the corresponding heat flux is $Ts_m A = (\rho + P)A$. Balancing the two gives us the entropy (or heat) balance condition $T\,dS/dt = s_m AT$, which becomes

$$\frac{H}{8\pi G}\frac{dA}{dt} = (\rho + P)A \,. \tag{12.35}$$

Using $A = 4\pi/H^2$, this gives the result

$$\dot{H} = -4\pi G(\rho + P) \,, \tag{12.36}$$

which is the correct Friedmann equation. Combining with the energy conservation for matter $\rho\, da^3 = -P\, da^3$, we immediately find that

$$\frac{3H^2}{8\pi G} = \rho + \text{constant} = \rho + \rho_\Lambda \,, \tag{12.37}$$

where ρ_Λ is the energy density of the cosmological constant (with $P_\Lambda = -\rho_\Lambda$) which arises in the form of an integration constant. We thus see that the entropy balance condition correctly reproduces the field equation – but with an arbitrary cosmological constant arising as the integration constant. This is obvious from the fact that, treated as a fluid, the entropy density $(s_\Lambda = (1/T)(\rho_\Lambda + P_\Lambda) = 0)$ vanishes for the cosmological constant. Thus, one can always add an arbitrary cosmological constant without affecting the entropy balance.

This is a general feature of the emergent paradigm and has important consequences for the cosmological constant problem. In the conventional approach, gravity is treated as a field which couples to the *energy density* of matter. The addition of a cosmological constant – or equivalently, shifting of the zero level of the energy – is not a symmetry of the theory and the field equations (and their solutions) change under such a shift. In the emergent perspective, it is the *entropy density* rather than the *energy density* which plays the crucial role. When the spacetime responds in a manner maintaining

entropy balance, it responds to the combination $\rho + P$ (or, more generally, to $T_{ab}n^a n^b$, where n^a is a null vector), which vanishes for the cosmological constant. In other words, shifting of the zero level of the energy is the symmetry of the theory in the emergent perspective and gravity does not couple to the cosmological constant. Alternatively, one can say that the restoration of this symmetry allows us to gauge away any cosmological constant, thereby setting it to zero. From this point of view, the vanishing of the bulk cosmological constant is a direct consequence of a symmetry in the theory. We will see later in Sect. 12.6 that the presence of a small cosmological constant or dark energy in the universe has to be thought of as a relic from quantum gravity when this symmetry is broken.

Deformations of Null Surfaces

The Noether current provides a nice, alternative description of the result that field equations become an entropy balance law on the null surface and connects it up with the surface term in the action functional. We will now briefly discuss this feature.

To any conserved current J^a, we can associate with an infinite family of vector fields q^l through the equation $J^c \equiv \nabla_l(\nabla^c q^l - \nabla^l q^c)$. (This is obvious if we think of q^a as the electromagnetic vector potential produced by the conserved current J^a; two vector fields q_a and $q_a + \partial_a \alpha$ belong to the same family and produce the same Noether potential and current.) With straightforward algebraic manipulation, we can now write this result as an *identity* satisfied by any conserved J^c

$$J^c = \nabla_l(\nabla^c q^l - \nabla^l q^c) = 2R^c_m q^m - \mathcal{V}^c\,, \tag{12.38}$$

with

$$\mathcal{V}^c \equiv g^{ik}\pounds_q \Gamma^c_{ik} - g^{ck}\pounds_q \Gamma^l_{kl} = g^{lm}\pounds_q N^c_{lm}\,, \tag{12.39}$$

where N^c_{lm} is the canonical momentum defined in (12.6). The two terms on the right-hand side of (12.38) arise from the variation of Einstein–Hilbert action under the diffeomorphism $x^i \to x^i + q^i$. So we can interpret the conservation of *any* current J^c as due to the diffeomorphism invariance of Einstein–Hilbert action under the spacetime deformation of a corresponding vector field q^a related to J^a! These ideas generalize to Lanczos–Lovelock theories. Given any conserved current J^a and an entropy tensor P^{abcd} it is possible to solve the equation $2P^{abcd}\nabla_b\nabla_c q_d = J^a$ and obtain an infinite set of q^as, again related to each other by a gauge transformation. Just as in the case of general relativity, one can now obtain an algebraic identity

$$J^c = 2\mathcal{R}^c_m q^m - \mathcal{V}^c\,, \tag{12.40}$$

where $\mathcal{V}^c \equiv 2P_a^{\ bcd}\pounds_q \Gamma^a_{bd}$. The conservation of this current now follows from invariance of the Lanczos–Lovelock Lagrangian (for which the chosen P^{abcd} is the entropy tensor) under the diffeomorphism induced by q^a.

Given a deformation field q^a we can separate its gradient $\nabla_a q_b$ into a symmetric and antisymmetric parts by $(1/2)\pounds_n g_{ab} = S_{ab} = (1/2)\nabla_{(a}q_{b)}$ and $(1/2)J_{ab} \equiv F_{ab} = (1/2)\nabla_{[a}q_{b]}$ in a standard manner. We will call a deformation q^a *isoentropic* at an event $\mathcal{P}$, if $J_{ab} = 0$ around that event and *Killing* if $S_{ab} = 0$ around that event. The most natural context in which isoentropic deformation arises is when we consider a deformation normal to a null surface; if $\phi(x) =$ constant describes a family of null surfaces, then its normal $q^a = \nabla^a \phi$ (locally) can be taken to be pure gradient since there is no unique normalization for a null vector. In this case, using (12.40), we can write the Lanczos–Lovelock field equations as

$$T_{ab}q^a q^b = \mathcal{V}^a q_a = q_c g^{lm}\pounds_q N^c_{lm}\,, \tag{12.41}$$

where the last relation holds for Einstein's theory. $T_{ab}q^a q^b$ (multiplied by $2\pi/\kappa$) is related to the matter entropy flux through the null surface, while $\mathcal{V}^a q_a$ is the gravitational entropy flux contribution. The fact the $\mathcal{V}^c$ term arises from the variation of the surface term in the action (which as we know is related to the entropy) shows the relationship between the two results. In Einstein's theory this contribution is related to the Lie derivative of the gravitational momentum under the deformation of the null surface.

12.4.2 The Avogadro Number of the Spacetime and Holographic Equipartition

Given the fact that spacetime appears to be hot, just like a body of gas, we can apply the Boltzmann paradigm (*If you can heat it, it has microstructure*) and study the nature of the microscopic degrees of freedom of the spacetime – exactly the way people studied gas dynamics *before* the atomic structure of matter was understood.

One key relation in such an approach is the equipartition law $\Delta E = (1/2)k_B T\,\Delta N$ relating the number density ΔN of microscopic degrees of freedom we need to

store an energy ΔE at temperature T. (This number is closely related to the Avogadro number of a gas, which was known even before people worked out what it was counting!). If gravity is the thermodynamic limit of the underlying statistical mechanics, describing the *atoms of spacetime*, we should be able to relate E and T of a given spacetime and determine the number density of microscopic degrees of freedom of the spacetime when everything is static. Remarkably enough, we can do this directly from the gravitational field equations [12.59–61]. Einstein's equations *imply* the equipartition law between the energy E in a volume V bounded by an equipotential surface ∂V and degrees of freedom on the surface

$$
\begin{aligned}
E &= \frac{1}{2}\int_{\partial V} \frac{\sqrt{\sigma}\,\mathrm{d}^2x}{L_P^2}\frac{\hbar}{c}\left\{\frac{Na^\mu n_\mu}{2\pi}\right\} \\
&\equiv \frac{1}{2}k_B\int_{\partial V} \mathrm{d}nT_{\mathrm{loc}}\,,
\end{aligned}
\tag{12.42}
$$

where $k_BT_{\mathrm{loc}} \equiv (\hbar/c)(Na^\mu n_\mu/2\pi)$ is the local acceleration temperature and $\Delta n \equiv \sqrt{\sigma}\,\mathrm{d}^2x/L_P^2$ with $\mathrm{d}A = \sqrt{\sigma}\,\mathrm{d}^2x$ being the proper surface area element. This allows us to read off the number density of microscopic degrees of freedom. We see that, unlike normal matter – for which the microscopic degrees of freedom scale in proportion to the volume and one would have obtained an integral over the volume of the form $\mathrm{d}V(\mathrm{d}n/\mathrm{d}V)$ – the degrees of freedom now scale in proportion to *area* of the boundary of the surface. In this sense, gravity is holographic. In Einstein's theory, the number density $(\mathrm{d}n/\mathrm{d}A) = 1/L_P^2$ is a constant with every Planck area contributing a single degree of freedom. The true importance of these results again rests on the fact that they remain valid for all Lanczos–Lovelock models with correct surface density of degrees of freedom [12.61].

Holographic Equipartition in the Newtonian Limit

Considering the importance of the above result for cosmology, we will provide [12.60] an elementary derivation of this result in the Newtonian limit of general relativity, to leading order in c^2. Consider a region of three-dimensional space V *bounded by an equipotential surface* ∂V, containing mass density $\rho(t,\boldsymbol{x})$ and producing a Newtonian gravitational field $\boldsymbol{g}$ through the Poisson equation $-\nabla\cdot\boldsymbol{g} \equiv \nabla^2\phi = 4\pi G\rho$. Integrating ρc^2 over the region V and using the Gauss law, we obtain

$$
\begin{aligned}
E = Mc^2 &= -\frac{c^2}{4\pi G}\int_V \mathrm{d}V\nabla\cdot\boldsymbol{g} \\
&= \frac{c^2}{4\pi G}\int_{\partial V} \mathrm{d}A(-\hat{\boldsymbol{n}}\cdot\boldsymbol{g})\,.
\end{aligned}
\tag{12.43}
$$

Since ∂V is an equipotential surface, $-\hat{\boldsymbol{n}}\cdot\boldsymbol{g} = g$ is the magnitude of the acceleration at any given point on the surface. Once again, introducing a $\hbar$ into this classical Newtonian law to bring in the Davies–Unruh temperature $k_BT = (\hbar/c)(g/2\pi)$ we obtain the result

$$
\begin{aligned}
E &= \frac{c^2}{4\pi G}\int_{\partial V}\mathrm{d}Ag = \int_{\partial V}\frac{\mathrm{d}A}{(G\hbar/c^3)}\frac{1}{2}\left(\frac{\hbar}{c}\frac{g}{2\pi}\right) \\
&= \int_{\partial V}\frac{\mathrm{d}A}{(G\hbar/c^3)}\left(\frac{1}{2}k_BT\right),
\end{aligned}
\tag{12.44}
$$

which is exactly the Newtonian limit of the holographic equipartition law in (12.42).

In the still simpler context of spherical symmetry, the integration over $\mathrm{d}A$ becomes multiplication by $4\pi R^2$, where R is the radius of the equipotential surface under consideration, and we can write the equipartition law as an equality between number of degrees of freedom in the bulk and surface

$$
N_{\mathrm{bulk}} = N_{\mathrm{sur}}\,, \tag{12.45}
$$

where

$$
\begin{aligned}
N_{\mathrm{bulk}} &\equiv \frac{E}{(1/2)k_BT}\,, \quad N_{\mathrm{sur}} = \frac{4\pi R^2}{L_P^2}\,, \\
E &= M(<R)c^2\,, \quad k_BT = \frac{\hbar}{c}\frac{GM}{2\pi R^2}\,.
\end{aligned}
\tag{12.46}
$$

In this form, we can think of $N_{\mathrm{bulk}} \equiv [E/(1/2)k_BT]$ as the degrees of freedom of the matter residing in the bulk and (12.46) can be thought of as providing the equality between the degrees of freedom in the bulk and the degrees of freedom on the boundary surface. We will call this *holographic equipartition*, which among other things, implies a quantization condition on the bulk energy contained inside any equipotential surface.

In the general relativistic case, the source of gravity is proportional to $\rho c^2 + 3P$ rather than ρ. In the nonrelativistic limit, ρc^2 will dominate over P and the equipartition law $E = (1/2)N_{\text{sur}}k_{\text{B}}T$ relates the rest mass energy Mc^2 to the surface degrees of freedom N_{sur}. If we instead decide to use the normal kinetic energy $E_{\text{kin}} = (1/2)Mv^2$ of the system (where $v = (GM/R)^{1/2}$ is the typical velocity determined through, say, the virial theorem $2E_{\text{kin}} + U_{\text{grav}} = 0$), then we have the result

$$\begin{aligned} E_{\text{kin}} &= \frac{v^2}{2c^2}E = \frac{v^2}{2c^2}\left(\frac{1}{2}N_{\text{sur}}k_{\text{B}}T\right) \\ &\equiv \frac{1}{2}N_{\text{eff}}k_{\text{B}}T\,, \end{aligned} \tag{12.47}$$

where

$$N_{\text{eff}} \equiv \frac{v^2}{2c^2}N_{\text{sur}} = 2\pi\frac{MRc}{\hbar} \tag{12.48}$$

can be thought of as the *effective* number of degrees of freedom which contributes to holographic equipartition with the kinetic energy of the self-gravitating system. In virial equilibrium, this kinetic energy is essentially $E_{\text{kin}} = (1/2)|U_g|$ and hence the gravitational potential energy inside an equipotential surface is also determined by N_{eff} by

$$\begin{aligned} |U_{\text{grav}}| &= \frac{1}{8\pi G}\int_V \text{d}V|\nabla\phi|^2 = 2E_{\text{kin}} \\ &= N_{\text{eff}}k_{\text{B}}T = 2\pi\frac{MRc}{\hbar}k_{\text{B}}T\,. \end{aligned} \tag{12.49}$$

We thus find that, for a nonrelativistic Newtonian system, the rest mass energy corresponds to $N_{\text{sur}} \propto (R^2/L_{\text{P}}^2)$ of surface degrees of freedom in holographic equipartition, while the kinetic energy and gravitational potential energy corresponds to the number of degrees of freedom $N_{\text{eff}} \propto MR$ which is smaller by a factor v^2/c^2. In the case of a black hole, $M \propto R$, making $MR \propto R^2$ leading to the equality of all these expressions. We will see later on that the difference $(N_{\text{sur}} - N_{\text{bulk}})$ plays a crucial role in cosmology.

The Approach to Holographic Equipartition

When the spacetime is not static, we do not expect the equipartition law to hold and the difference between the bulk and surface degrees of freedom will drive the dynamical evolution. To see this, consider a spacetime in which we have introduced the usual $(1+3)$ split with the normals to $t =$ constant surfaces being u^a, which we can take to be the four-velocities of a congruence of observers. Let $a^i \equiv u^j\nabla_j u^i$ be the acceleration of the congruence and $K_{ij} = -\nabla_i u_j - u_i a_j$ be the extrinsic curvature tensor. We then have the identity

$$\begin{aligned} R_{ab}u^a u^b &= \nabla_i(Ku^i + a^i) + K^2 - K_{ab}K^{ab} \\ &= u^a\nabla_a K + \nabla_i a^i - K_{ij}K^{ij}\,. \end{aligned} \tag{12.50}$$

When the spacetime is static, we can choose a natural coordinate system with $K_{ij} = 0$ so that the above equation reduces to $\nabla_i a^i = R_{ab}u^a u^b$. Using the field equations to write $R_{ab}u^a u^b = 8\pi\bar{T}_{ab}u^a u^b$ (where $\bar{T}_{ab} = T_{ab} - (1/2)g_{ab}T$) and integrating $\nabla_i a^i = 8\pi\bar{T}_{ab}u^a u^b$ over a region of space, we can immediately obtain the equipartition law discussed in Sect. 12.4.1.

In a general spacetime, if we choose a local gauge with $N_\alpha = 0$, $u_i = -N\delta_i^0$ (so that $a^0 = 0$), then (12.50) can be reduced to the form

$$D_\mu(Na^\mu) = 4\pi\rho_{\text{Komar}} + N\left(K^\alpha_\beta K^\beta_\alpha - \dot{K}\right), \tag{12.51}$$

where

$$\begin{aligned} \rho_{\text{Komar}} &\equiv 2N\bar{T}_{ab}u^a u^b\,; \\ \dot{K} &\equiv \text{d}K/\text{d}\tau \equiv u^a\nabla_a K\,. \end{aligned} \tag{12.52}$$

Integrating this relation over a region of space, we can express the departure from equipartition, as seen by observers following this congruence as

$$\begin{aligned} E - \frac{1}{2}\int_{\partial V} k_{\text{B}}T_{\text{loc}}\,\text{d}n &= \frac{1}{4\pi}\int_V \text{d}^3x\sqrt{h}N \\ &\quad \times\left(\dot{K} - K^\alpha_\beta K^\beta_\alpha\right). \end{aligned} \tag{12.53}$$

This is an exact equation which can be used to study the evolution of the geometry in terms of the departure from equipartition for both finite and cosmological systems.

These results suggest that one should be able to think of gravitational dynamics from a completely different perspective closer in spirit to the manner in which we view the bulk properties of matter like elasticity or fluid dynamics. We will now explore this aspect. For further developments in this direction see [12.62].

12.5 Gravity from an Alternative Perspective

If we take this point of view seriously, then the deformations of spacetime $(\bar{x}^i - x^i) \equiv q^i$ associated with a vector field q^i are analogous to deformations of a solid in the study of elasticity. By and large, such a spacetime deformation is not of much consequence except when we consider the deformations of null surfaces. As we described earlier, any null surfaces can be thought of as acting as a local Rindler horizon to a suitable set of observers. The deformation of a local patch of a null surface will change the amount of information accessible to the local Rindler observer. Therefore, such an observer will associate a certain amount of entropy density with the deformation of a null patch with normal n^a. We might hope that extremizing the sum of gravitational and matter entropy associated with *all* null vector fields *simultaneously* could then lead to the equations obeyed by the background metric.

Conceptually, this idea is very similar to the manner in which we determine the influence of gravity on other matter fields. If we fill the spacetime with freely falling observers and insist that normal laws of special relativity should hold for all these observers simultaneously, we can arrive at the generally covariant versions of equations obeyed by matter in an arbitrary metric. This, in turn, allows us to determine the influence of gravity on matter fields, thereby fixing the *kinematics* of gravity. To determine the *dynamics*, we play the same game but now by filling the spacetime with local Rindler observers. Insisting that the local thermodynamics should lead to the extremum of an entropy functional associated with every null vector in the spacetime, we will obtain a set of equations that will determine the background spacetime.

There is no a priori assurance that such a program will succeed and hence it is yet another surprise that one can actually achieve this. Let us associate with every null vector field $n^a(x)$ in the spacetime a thermodynamic potential $\Im(n^a)$ (say, entropy) which is quadratic in n^a and is given by

$$\begin{aligned}\Im[n^a] &= \Im_{\text{grav}}[n^a] + \Im_{\text{matt}}[n^a] \\ &\equiv -\left(4P_{ab}^{cd}\nabla_c n^a \nabla_d n^b - T_{ab}n^a n^b\right),\end{aligned} \tag{12.54}$$

where P_{ab}^{cd} and T_{ab} are two tensors that play a role analogous to elastic constants in the theory of elastic deformations. If we extremize this expression with respect to n^a, we will normally obtain a differential equation for n^a involving its second derivatives. We, however, want to demand that the extremum holds for all n^a, thereby constraining the *background* geometry. Further, our insistence on the strictly local description of null-surface thermodynamics translates into the demand that the Euler derivative of the functional $\Im(n^a)$ should not contain any derivatives of n^a.

It is indeed possible to satisfy all these conditions by the following choice. We take P_{ab}^{cd} to be a tensor having the symmetries of a curvature tensor and being divergence-free in all its indices; we take T_{ab} to be a divergence-free symmetric tensor. (The conditions $\nabla_a P^{ab}_{cd} = 0, \nabla_a T^a_b = 0$ can also be thought of as a generalization of the notion of *constancy* of elastic constants of spacetime [12.63].) Once we obtain the field equations we can read T_{ab} as the matter energy-momentum tensor; the notation anticipates this result. We also know that the P^{abcd} with the assigned properties can be expressed as $P_{ab}^{cd} = \partial L/\partial R^{ab}_{cd}$, where L is the Lanczos–Lovelock Lagrangian and R_{abcd} is the curvature tensor [12.1, 2]. This choice in (12.54) will also ensure that the equations resulting from the entropy extremization do not contain any derivative of the metric which is higher than second order.

We now demand that $\delta\Im/\delta n^a = 0$ for the variation of all null vectors n^a with the condition $n_a n^a = 0$ imposed by adding a Lagrange multiplier function $\lambda(x) g_{ab} n^a n^b$ to $\Im[n^a]$. An elementary calculation and use of generalized Bianchi identity and the condition $\nabla_a T^a_b = 0$ leads us to [12.1, 2, 64–66] the following equations for background geometry

$$\mathcal{G}^a_b = \mathcal{R}^a_b - \frac{1}{2}\delta^a_b L = \frac{1}{2}T^a_b + \Lambda\delta^a_b\,, \tag{12.55}$$

where Λ is an integration constant. These are precisely the field equations for gravity in a theory with Lanczos–Lovelock Lagrangian L with an undetermined cosmological constant Λ, which arises as an integration constant.

The thermodynamical potential corresponding to the density $\Im$ can be obtained by integrating the density $\Im[n^a]$ over a region of space or a surface, etc., depending on the context. The matter part of the $\Im$ is proportional to $T_{ab}n^a n^b$, which will pick out the contribution $(\rho + p)$ for an ideal fluid, which is the enthalpy density. If multiplied by $\beta = 1/T$, this reduces to the entropy density because of the Gibbs–Duhem relation. When the multiplication by β can be reinterpreted in terms of integration over $(0, \beta)$ of the time coordinate (in the Euclidean version of the LRF), the corresponding potential can be interpreted as entropy and the

integral over space coordinates alone can be interpreted as the rate of generation of entropy. (This was the interpretation provided in the earlier works [12.1, 2, 64–66], but the result is independent of this interpretation as long as suitable boundary conditions can be imposed.) One can also think of $\mathfrak{J}[n^a]$ as an effective Lagrangian for a set of collective variables n^a describing the deformations of null surfaces.

The gravitational entropy density in (12.54) can be expressed in terms of the Killing and isoentropic deformations introduced in the section *Deformations of Null Surfaces*

$$\begin{aligned}-4P^{ab}_{cd}\nabla_a q^c \nabla_b q^d &= 4P^{bijd}S_{ij}S_{bd} \\ &\quad - 2P^{abcd}F_{ab}F_{cd} \\ &= P^{bijd}(\pounds_q g_{ij})(\pounds_q g_{bd}) \\ &\quad - \frac{1}{2}P^{abcd}J_{ab}J_{cd}\,. \end{aligned} \tag{12.56}$$

The second equation shows that the gravitational entropy density has two parts: one coming from the square of the Noether potential (which vanish for isoentropic deformations) and another which depends on the change in the metric under the deformation (which will vanish for the Killing deformations). When q_j is a pure gradient, J_{ab} will vanish and one can identify the first term with a structure like $\mathrm{Tr}(K^2) - (\mathrm{Tr}K)^2$. On the other hand, when q_a is a local Killing vector, the contribution from S_{ij} to the entropy density vanishes and we find that the entropy density is just the square of the antisymmetric potential J_{ab}. For a general null vector, both the terms contribute to the entropy density. Variation of entropy density with respect to either of the two contributions (after adding suitable Lagrange multiplier to ensure vanishing of the other term) will lead to the gravitational field equations.

In this approach, there arise several new features worth mentioning.

First, we find [12.64, 65] that the extremum value of the thermodynamic potential, when computed on-shell for a solution with static horizon, leads to the Wald entropy. This is a nontrivial consistency check on the approach because it was not designed to reproduce the Wald entropy. When the field equations hold, the total entropy of a region $\mathcal{V}$ resides on its boundary $\partial\mathcal{V}$, which is yet another illustration of the holographic nature of gravity.

Second, in the semi-classical limit, one can show [12.67] that the gravitational (Wald) entropy is quantized with S_{grav} [on-shell] $= 2\pi n$. In the lowest order Lanczos–Lovelock theory, the entropy is proportional to area and this result leads to area quantization. More generally, it is the gravitational entropy that is quantized. The law of equipartition for the surface degrees of freedom is closely related to this entropy quantization.

Third, the entropy functional in (12.54) is invariant under the shift $T_{ab} \to T_{ab} + \rho_0 g_{ab}$, which shifts the zero of the energy density. This symmetry allows any low energy cosmological constant, appearing as a parameter in the variational principle, to be gauged away thereby alleviating the cosmological constant problem to a great extent [12.66, 68, 69]. As far as we know, *this is the only way in which one can make gravity immune to the zero point level of energy density.* It is again interesting that our approach leads to this result in a natural fashion even though it is not designed for this purpose. This works because the cosmological constant, treated as an ideal fluid, has zero entropy because $\rho + p = 0$ and thus cannot affect gravitational dynamics in this perspective in which gravity responds to the entropy density rather than energy density.

Fourth, the *algebraic* reason for the whole idea to work is the easily proved identity

$$\begin{aligned}4P^{cd}_{ab}\nabla_c n^a \nabla_d n^b &= 2\mathcal{R}_{ab}n^a n^b \\ &\quad + \nabla_c\left[4P^{cd}_{ab}n^a\nabla_d n^b\right], \end{aligned} \tag{12.57}$$

which shows that, except for a boundary term, we are extremizing the integral of $(2R_{ab} - T_{ab})n^a n^b$ with respect to n^a subject to the constraint $n_a n^a = 0$. The algebra is trivial but not the underlying concept. In fact, if we ignore the total divergence term in (12.57) then we can express the total entropy in a spacetime region as

$$S = \int_{\partial\partial\mathcal{V}} \mathrm{d}^{D-2}\Sigma_{ab}J^{ab} + \int_{\partial\mathcal{V}} \mathrm{d}^{D-1}\Sigma_a \mathcal{V}^a\,. \tag{12.58}$$

The first term is the contribution from the Noether potential on a surface of codimension 2 (which vanishes if $n_a = \nabla_a\phi$), while the second term gives the contribution from the variation of the surface term in the action. In writing this expression, we have assumed suitable boundary conditions to ignore contributions from other boundaries. As explained before, the contribution from the Noether potential vanishes for isoentropic deformations and the contribution from the action vanishes for Killing deformations.

Fifth, the gravitational entropy density – which is the term in the integrand $\Im_{\text{grav}} \propto (-P^{cd}_{ab}\nabla_c n^a \nabla_d n^b)$ in (12.54) – also obeys the relation

$$\frac{\partial \Im_{\text{grav}}}{\partial (\nabla_c n^a)} = -8\left(-P^{cd}_{ab}\nabla_d n^b\right) = \frac{1}{4\pi}\left(\nabla_a n^c - \delta^c_a \nabla_i n^i\right), \tag{12.59}$$

where the second relation is for Einstein's theory. This term is analogous to the more familiar object $t^c_a = K^c_a - \delta^c_a K$ (where K_{ab} is the extrinsic curvature) that arises in the $(1+3)$ separation of Einstein's equations. (More precisely, the projection to three-space leads to t^c_a.) This term has the interpretation as the canonical momentum conjugate to the spatial metric in (1+3) context, and (12.59) shows that the entropy density leads to a similar structure. That is, the canonical momentum conjugate to the metric in the conventional approach and the momentum conjugate to n^a in $\Im_{\text{grav}}$ are essentially the same. For further developments in this direction see [12.62].

12.6 Emergence of Cosmic Space

In the discussion of the emergent paradigm so far, we argued that the *field equations are emergent* while assuming the existence of a spacetime manifold, metric, curvature, etc., as given structures. In this approach, we interpret the field equations as certain consistency conditions obeyed by the background spacetime. A more ambitious project will be to give meaning to the concept that the *spacetime itself is an emergent structure*. The idea is to build up the spacetime itself from some underlying pregeometric variables, along the lines we obtain macroscopic variables like density, temperature, etc., from atomic properties of matter. Unfortunately, it is not easy to give this idea a rigorous mathematical expression, consistent with what we know already know about space and time. In attempting this, we run into (at least) two key difficulties.

The first issue has to do with the role played by time, which is quite different from the role played by space in the entire description of physics. It is very difficult conceptually to treat time as being emergent from some pregeometric variable if it has to play the standard role of a parameter that describes the evolution of the dynamical variables. It seems necessary to treat time differently from space, which runs counter to the spirit of general covariance.

The second issue has to do with space around *finite* gravitating systems, like the Earth, Sun, Milky Way, etc. It is incorrect to argue that space is emergent around such *finite* gravitating systems because direct experience tells us that space around them is preexisting. So any emergent description of the gravitational fields of *finite systems* has to work with space as a given entity – along the lines we described in the previous sections. Thus, when we deal with *finite* gravitating systems, without assigning any special status to a time variable, it seems impossible to come up with a conceptually consistent formulation for the idea that *spacetime itself is an emergent structure*.

What is remarkable is the fact that both these difficulties disappear [12.70] when we consider spacetime in the cosmological context! Observations show that there is, indeed, a special choice of time variable available in our universe, which is the proper time of the geodesic observers who see the CMBR as homogeneous and isotropic. This fact justifies treating time differently from space in (and *only* in) the context of cosmology. Further, the spatial expansion of the universe can certainly be thought of as equivalent to the emergence of space as the cosmic time flows forward. All this suggests that we may be able to make concrete the idea that *cosmic space is emergent as cosmic time progresses* in a well-defined manner in the context of cosmology. We will now describe how it works.

Once we assume that the expansion of the universe is equivalent to emergence of space, we need to ask *why* this happens. In the more conservative approach described in earlier sections, the dynamics of spacetime is governed by gravitational field equations and we can obtain the expanding universe as a *special solution* to these equations. However, when we want to treat space itself as being emergent, we cannot start with gravitational field equations and need to treat cosmic evolution as more fundamental.

The holographic principle suggests a deep relationship between the number of degrees of freedom residing in a bulk region of space and the number of degrees of freedom on the boundary of this region. To see *why* cosmic space emerges – or, equivalently, why the universe is expanding – we will use a specific version of

holographic principle. To motivate this, let us consider a pure de Sitter universe with a Hubble constant H. Such a universe obeys the holographic principle in the form

$$N_{\rm sur} = N_{\rm bulk} \,. \tag{12.60}$$

Here the $N_{\rm sur}$ is the number of degrees of freedom attributed to a spherical surface of Hubble radius H^{-1} and is given by

$$N_{\rm sur} = \frac{4\pi}{L_{\rm P}^2 H^2}\,, \tag{12.61}$$

if we attribute one degree of freedom per Planck area of the surface. $N_{\rm bulk} = |E|/[(1/2)k_{\rm B}T]$ is the *effective* number of degrees of freedom which are in equipartition at the horizon temperature $k_{\rm B}T = (H/2\pi)$, with $|E|$ being the Komar energy $|(\rho + 3P)|V$ contained inside the Hubble volume $V = (4\pi/3H^3)$. So

$$N_{\rm bulk} = -\frac{E}{(1/2)k_{\rm B}T} = -\frac{2(\rho+3P)V}{k_{\rm B}T}\,. \tag{12.62}$$

For a pure de Sitter universe with $P = -\rho$, our (12.60) reduces to $H^2 = 8\pi L_{\rm P}^2\rho/3$, which is the standard result. Note that $(\rho + 3P)$ is the proper Komar energy density, while $V = 4\pi/3H^3$ is the *proper* volume of the Hubble sphere. The corresponding *comoving* expressions will differ by a^3 factors in both, which will cancel out leading to the same expression for E.

This result is consistent with the equipartition law described earlier in Sect. 12.4.2 in which we obtained the result $|E| = (1/2)N_{\rm sur}k_{\rm B}T$ (which is, of course, the same as (12.60)) *as a consequence of* gravitational field equations in static spacetimes. Here, we do not assume any field equations but will consider the relation $|E|/(1/2)k_{\rm B}T = N_{\rm sur}$ as fundamental. The (12.60) represents the *holographic equipartition* and relates the effective degrees of freedom residing in the bulk, determined by the equipartition condition, to the degrees of freedom on the boundary surface. The dynamics of the pure de Sitter universe can thus be obtained directly from the holographic equipartition condition, taken as the starting point.

Our universe, of course, is not pure de Sitter but is evolving towards an asymptotically de Sitter phase. It is, therefore, natural to think of the current accelerated expansion of the universe as an evolution towards holographic equipartition. Treating the expansion of the universe as conceptually equivalent to the emergence of space we conclude that the emergence of space itself is being driven towards holographic equipartition. Then we expect the law governing the emergence of space to relate the availability of greater and greater volumes of space to the departure from holographic equipartition given by the difference $(N_{\rm sur} - N_{\rm bulk})$. The simplest (and the most natural) form of such a law will be

$$\Delta V = \Delta t (N_{\rm sur} - N_{\rm bulk})\,, \tag{12.63}$$

where V is the Hubble volume in Planck units and t is the cosmic time in Planck units. Our arguments suggest that $(\Delta V/\Delta t)$ will be some function of $(N_{\rm sur} - N_{\rm bulk})$, which vanishes when the latter does. Then, (12.63) represents the Taylor series expansion of this function truncated at the first order. We will now elevate this relation to the status of a postulate which governs the emergence of the space (or, equivalently, the expansion of the universe) and show that it is equivalent to the standard Friedmann equation. Reintroducing the Planck scale and setting $(\Delta V/\Delta t) = {\rm d}V/{\rm d}t$, this equation becomes

$$\frac{{\rm d}V}{{\rm d}t} = L_{\rm P}^2 (N_{\rm sur} - N_{\rm bulk})\,. \tag{12.64}$$

Substituting $V = (4\pi/3H^3)$, $N_{\rm sur} = \left(4\pi/L_{\rm P}^2H^2\right)$, $k_{\rm B}T = H/2\pi$ and using $N_{\rm bulk}$ in (12.62), we find that the left-hand side of (12.64) is proportional to ${\rm d}V/{\rm d}t \propto (-\dot{H}/H^4)$, while the first term on the right-hand side gives $N_{\rm sur} \propto (1/H^2)$. Combining these two terms and using $\dot{H} + H^2 = \ddot{a}/a$, it is easy to show that this equation simplifies to the relation

$$\frac{\ddot{a}}{a} = -\frac{4\pi L_{\rm P}^2}{3}(\rho + 3P)\,, \tag{12.65}$$

which is the standard dynamical equation for the Friedmann model. The condition $\nabla_a T^a_b = 0$ for matter gives the standard result ${\rm d}(\rho a^3) = -P{\rm d}a^3$. Using this, (12.65), and the de Sitter boundary condition at late times, one gets back the standard accelerating universe scenario. Thus, we can describe the evolution of the accelerating universe entirely in terms of the concept of holographic equipartition.

Let us next consider the full evolution of the universe, consisting of both the decelerating and accelerating phases. The definition of $N_{\rm bulk}$ in (12.62) makes sense only in the accelerating phase of the universe where $(\rho + 3P) < 0$ so as to ensure $N_{\rm bulk} > 0$. For normal matter, we would like to use (12.62) without the

negative sign. This is easily taken care of by using appropriate signs for the two different cases and writing

$$\frac{dV}{dt} = L_P^2(N_{\rm sur} - \epsilon N_{\rm bulk}) \,, \tag{12.66}$$

with the definition

$$N_{\rm bulk} = -\epsilon \frac{2(\rho + 3P)V}{k_B T} \,. \tag{12.67}$$

Here $\epsilon = +1$ for matter with $(\rho + 3P) < 0$ and $\epsilon = -1$ for matter with $(\rho + 3P) > 0$. (We use the sign convention such that we maintain the form of (12.63) for the accelerating phase of the universe. One could have, of course, used the opposite convention for ϵ and omitted the minus sign in (12.67).) Because only the combination $+\epsilon^2(\rho + 3P) \equiv (\rho + 3P)$ occurs in (dV/dt), the derivation of (12.65) remains unaffected and we also maintain $N_{\rm bulk} > 0$ in all situations (Fig. 12.1). We can understand (12.66) better if we separate out the matter component, which causes deceleration, from the dark energy, which causes acceleration. For the sake of simplicity, we will assume that the universe has just two components (pressureless matter and dark energy) with $(\rho + 3P) > 0$ for matter and $(\rho + 3P) < 0$ for dark energy. In that case, (12.66) can be expressed in an equivalent form as

$$\frac{dV}{dt} = L_P^2(N_{\rm sur} + N_m - N_{de}) \,, \tag{12.68}$$

where all the three degrees of freedom, $N_{\rm sur}$, and N_m, N_{de}, are positive (as they should be) with $(N_m - N_{de}) = (2V/k_B T)(\rho + 3P)_{\rm tot}$. We now see that the condition of holographic equipartition with the emergence of space coming to an end $(dV/dt \to 0)$ asymptotically, can be satisfied only if we have a component in the universe with $(\rho + 3P) < 0$. In other words, *the existence of a cosmological constant in the universe is required for asymptotic holographic equipartition.*

One can, of course, obtain (12.66) from the general result in (12.53). In the Friedmann universe, the natural observers are the geodesic observers for whom $a^i = 0$. For the geodesic observers, (12.50) reduces to

$$u^a \nabla_a K \equiv \dot{K} = K_{ij}K^{ij} + 8\pi \bar{T}_{ab}u^a u^b \,. \tag{12.69}$$

Further, in the Friedmann universe, $K^\alpha_\beta = -H\delta^\alpha_\beta$, giving $\dot{K} = -3\dot{H}; K_{ij}K^{ij} = 3H^2$. Using these values and dividing (12.69) throughout by H^4, it is easy to reduce it to (12.64). We can see that the surface degrees of freedom actually arise from a term of the kind $K_{ij}K^{ij}/K^4$,

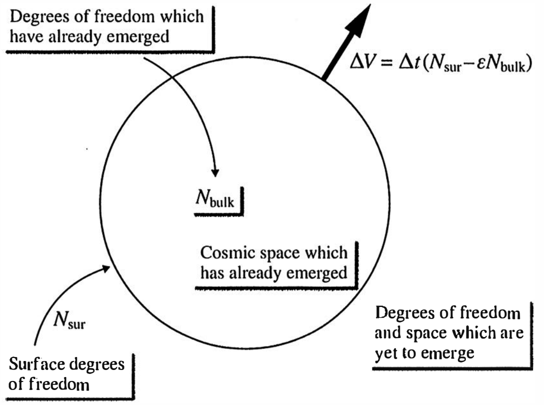


Fig. 12.1 This figure illustrates the ideas described in this section schematically. The shaded region represents the cosmic space that has already emerged by the time t, along with the surface degrees of freedom ($N_{\rm sur}$), which reside on the surface of the Hubble sphere and the bulk degrees of freedom ($N_{\rm bulk}$) that have reached equipartition with the Hubble temperature $k_B T = H/2\pi$. At this moment of time, the universe has not yet achieved the holographic equipartition. The holographic discrepancy $(N_{\rm sur} - \epsilon N_{\rm bulk})$ between these two drives the further emergence of cosmic space, measured by the increase in the volume of the Hubble sphere with respect to cosmic time, as indicated by the equation. Remarkably enough, this equation correctly reproduces the entire cosmic evolution

when one interprets $1/K$ as the relevant radius. However, our goal here is to think of (12.66) as the starting point of description.

Treating the Hubble radius $H^{-1}(t)$ as the boundary of cosmic space should not be confused with the causal limitation imposed by light propagation in the universe. If the Hubble radius at time t_1, say, is $H^{-1}(t_1)$, we assume that space of size $H^{-1}(t_1)$ can be thought of as having emerged for all $t \le t_1$. This is in spite of the fact that, at an earlier time $t < t_1$, the Hubble radius $H^{-1}(t)$ could have been significantly smaller. This is necessary for consistent interpretation of cosmological observations. For example, CMBR (cosmic microwave background radiation) observations allow us to probe, on the $z = z_{\rm rec} \approx 10^3$ surface, length scales that are larger than the Hubble radius $H^{-1}(t_{\rm rec})$ at $z = z_{\rm rec}$. So, as far as observations made today are concerned, we should assume that the size of the space that has emerged is the present Hubble radius, H_0^{-1}, rather than the instantaneous Hubble radius corresponding to the redshift of the epoch from which photons are

received. In this sense, the emergence of space from pregeometric variables may seem to be acausal, but it is completely consistent with what we know about the universe today.

As noted before, at the largest scales, our universe breaks Lorentz symmetry in the sense that our absolute motion with respect to the CMBR (treated as cosmic aether) can be (and has been) measured. Usually, one bypasses this embarrassing observational fact by saying that, while the known *laws of physics* exhibit a large symmetry (viz. general covariance) a *specific solution* of the gravitational field equations (like FRW models) need not exhibit the symmetry. This is perfectly correct in the conventional picture, where we think of gravitational field equations as fundamental and cosmology as described a special solution to these equations. In the description given above, however, we describe the cosmological evolution from a different principle and not as a special solution to field equations. We then start with the dynamics at largest scales and move to smaller scales. The matter and spacetime degrees of freedom emerge in the process and a higher degree of symmetry (general covariance) is realized at smaller scales while a preferred time coordinate continues to exist at largest scales. When the Hubble radius is endowed with a horizon temperature T, we can treat the bulk degrees of freedom which have *already emerged* – along with the space – as though they are inside a microwave oven with the temperature set to the surface value. Because *these* degrees of freedom account for an energy E, it follows that $E/(1/2)k_{\text{B}}T$ is, indeed, the correct count for *effective* N_{bulk}. (This temperature T and N_{bulk} should not be confused with the normal kinetic temperature of matter in the bulk and the standard degrees of freedom we associate with matter. It is more appropriate to think of these degrees of freedom as those which have already emerged, along with space, from some pregeometric variables.) The emergence of cosmic space itself is driven by the holographic discrepancy $(N_{\text{sur}}+N_m-N_{de})$ between the surface and bulk degrees of freedom, where N_m is contributed by normal matter with $(\rho+3P)>0$ and N_{de} is contributed by the cosmological constant with all the degrees of freedom being counted positive. In the absence of N_{de}, this expression can never be zero and holographic equipartition cannot be achieved. In the presence of the cosmological constant, the emergence of space will soon lead to N_{de} dominating over N_m when the universe undergoes accelerated expansion. Asymptotically, N_{de} will approach N_{sur} and the rate of emergence of space, $\mathrm{d}V/\mathrm{d}t$, will tend to zero allowing the cosmos to find its peace.

The study of the evolution of the universe using (12.63) is conceptually quite different from treating the expanding universe as a specific solution of gravitational field equations. The simplicity of (12.63) is quite striking and it is remarkable that the standard expansion of the universe can be reinterpreted as an evolution towards holographic equipartition. *If the underlying ideas are not correct, we need to explain why (12.63) holds in our universe!* The simplicity of (12.63) itself suggests proper choices for various physical quantities. For example, we have assumed that the relevant temperature for obtaining N_{bulk} is given by $T=H/2\pi$ even when H is time-dependent. There is some amount of controversy in the literature regarding the correct choice for this temperature. One can obtain equations similar to (12.63) with other definitions of the temperature but none of the other choices leads to equations with the compelling naturalness of (12.63). The same is true as regards the volume element V, which we have taken as the Hubble volume; other choices lead to equations which have no simple interpretation.

It should be noted that (12.63) is parameter free when expressed in Planck units and can be given a simple combinatorial interpretation. If we think of time evolution in steps of Planck time ($t=t_n, n=1,2,\ldots$) and the volume of the space which has emerged by the n-th step as V_n, then (12.63) tells us that

$$V_{n+1}=V_n+(N_{\text{sur}}-\epsilon N_{\text{bulk}})\,, \tag{12.70}$$

which is just an algorithmic procedure in integers. When we understand the pregeometric variables better, we may be able to interpret (12.63) purely in combinatorial terms. An immediate consequence of the discretized version in (12.70) is that we expect significant departures from conventional evolution when the relevant degrees of freedom are of the order of unity. Modifications of this equation will help us to study the evolution of the universe close to the big bang in a quantum cosmological setting when the degrees of freedom are of order unity. However, we have now bypassed the usual complications related to the time coordinate. Postulating suitable corrections to the *bit dynamics* in (12.70) may provide an alternate way of tackling the singularity problem of classical cosmology.

12.7 A Principle to Determine the Value of the Cosmological Constant

The idea of holographic equipartition, described in the last section demands a cosmological constant but does not determine its numerical value. We will now describe how it can be done in a unified picture for the evolution of the universe.

The current observations indicate that the radiation (and matter) dominated epoch of the universe is sandwiched between two asymptotic de Sitter epochs of expansion, usually identified with the inflationary era and the epoch of late time acceleration. These two de Sitter phases are characterized by two length scales L_{UV} and L_{IR} corresponding to the respective Hubble radii. Given the fact that de Sitter geometry is invariant under time translation, we can interpret the radiation (and matter) dominated phase of the universe as a transition state which connects the two *equilibrium* (steady) states of the geometry in which $N_{sur} = N_{bulk}$. It seems natural to set the length scale L_{UV} of the initial de Sitter phase to be the Planck length L_P and the length scale of the accelerating phase L_{IR} to be $(3/\Lambda)^{1/2}$, where Λ is the cosmological constant. The ratio of these two length scales is conveniently expressed in terms of the dimensionless number

$$\Lambda L_P^2 \approx 10^{-122} \approx 3 \times e^{-281} , \qquad (12.71)$$

where the numerical value is determined by observations (with $\Omega_{DE} \approx 0.7$ and $h \approx 0.7$).

An important question in theoretical physics is to determine this numerical value from first principles. For an excellent theoretical review, see [12.71]; for a classification of approaches to cosmological constant problem, see, e.g., [12.72].

If a fundamental principle can be found which allows the determination of this number, then all the conventional difficulties associated with the cosmological constant (e.g., why is it fine-tuned, why does it dominate the universe now, etc.) will vanish. We will describe a principle, closely related to the ideas of holographic equipartition, which allows us to express this number in the form

$$\Lambda L_P^2 = 3\exp(-24\pi^2\mu) , \qquad (12.72)$$

where μ is a number of order unity and, in principle, calculable. *Observations match with the above expression when $\mu \approx 1.18$, which is extremely encouraging.* Even for the *natural* choice of $\mu = 1$, (12.72) predicts $\log\left(\Lambda L_P^2/3\right) = -103$ compared to the observed value of -122. We do not know of any other attempt that could express the number ΛL_P^2 essentially in terms of $e, \pi, \ldots$ etc. and get this close to the observed value!

Let us now describe how (12.72) is obtained. In Fig. 12.2 we have shown the relevant length scales involved with our universe. The thick red line ADBC denotes the Hubble radius $H^{-1}(a) = (\dot{a}/a)^{-1}$, which is constant asymptotically during the early inflationary epoch ($a < a_F$) and late time accelerating phase ($a > a_\Lambda$). In the intermediate epoch ($a_F < a < a_\Lambda$), the Hubble radius DB grows as a^2 if the universe is radiation dominated. (To be precise, there is a regime close to B when the universe becomes matter dominated which we have ignored for the moment. Note that this phase lasts only for about 4 decades while the radiation dominated phase lasts for about 24 decades.)

In principle, one can extend the asymptotic de Sitter phases (which represent the universe in steady, time translation invariant, state) into the past and future as long as one wants. However, there are two natural boundaries to these de Sitter phases (for a detailed discussion, see [12.3, 66, 73]). The boundary A (at $a = a_I$) in the inflationary phase is determined as follows. We know that cosmologically relevant modes exit the Hubble radius during the inflationary phase and then re-enter the Hubble radius during the radiation (and matter) dominated phase. (The proper wavelengths of

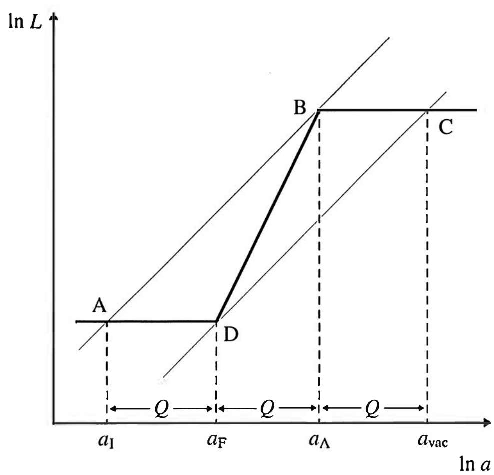


Fig. 12.2 The relevant length scales in a universe characterized by a radiation dominated epoch sandwiched between two de Sitter phases. See text for details

these modes grow linearly with a and will be lines of unit slope in Fig. 12.2.) *All* the modes which exit the Hubble radius during the inflationary phase will re-enter the Hubble radius later on if there is *no* late time accelerating phase for the universe. However, in the presence of late time acceleration, there is a mode with a critical wavelength (shown by the line AB, increasing linearly with a) which just skirts entering the Hubble radius. Given the points D and B, the point A is determined by drawing a unit slope line through B. Elementary geometry tells us that the universe expands by the same factor $Q \equiv (a_F/a_I) = (a_\Lambda/a_F)$ during AD and DB.

The natural boundary C (at $a = a_{vac}$) in the late time acceleration phase is determined by a different consideration. We know that the de Sitter phase is associated [12.13, 74] with the Gibbons–Hawking temperature $T_{dS} = (H/2\pi)$. As the universe expands, the CMBR temperature will keep decreasing $T_{cmb}(a) \propto (1/a)$ and will eventually fall below the de Sitter temperature T_{dS} after which the temperature of the universe will be essentially dominated by the de Sitter vacuum noise. This determines the point C through the condition $T_{cmb}(a_{vac}) = T_{dS}$. If the initial de Sitter phase is characterized by the Planck length (i. e., $H^{-1} = L_P$), then it is natural to assume that the reheating temperature at the end of Planck scale inflation is given by the de Sitter temperature $T_P = 1/(2\pi L_P)$. In that case, it is again elementary to show that the unit slope line drawn from C passes through D. (As an aside, note that we do *not* determine C by drawing a unit slope line from D. Instead, the point C is determined by the condition $T_{dS} = T_{cmb}$. A unit slope line drawn from C will not, in general, pass through the point D which is *independently* specified as the end point of inflation. It passes through D in this specific scenario because we have ignored the matter dominated phase and taken the reheating temperature to be set by Planck scale. However, even in a more realistic scenario, the three phases of expansion last for an approximately equal number of decades; for details, see, [12.66, 73].) So the relevant part of late time acceleration also lasts for an expansion by factor $Q = (a_{vac}/a_\Lambda)$.

In such a simple scenario, all the relevant physics is contained within the *cosmic parallelogram* ADCB with the universe expanding by the *same* factor Q during each of the three epochs. Given the fact that $H^{-1} \propto a^2$ during the radiation dominated phase DB, we can relate the Hubble radius $H_\Lambda^{-1} = (3/\Lambda)^{1/2}$ at B with the Hubble radius $H_P^{-1} = L_P$ at D by $(H_P/H_\Lambda) = (a_\Lambda/a_F)^2 = Q^2$. That is, $(H_\Lambda/H_P)^2 = (1/3)(\Lambda L_P^2) = Q^{-4}$, which allows us to relate the numerical value of the cosmological constant Λ to Q by

$$\Lambda L_P^2 = 3Q^{-4} . \tag{12.73}$$

The value of ΛL_P^2 is fixed, if we have a physical principle to determine Q, which is the factor by which the universe expands in each of these three phases. We will now describe such a physical principle.

To do this let us consider the modes (specified by comoving wave vectors k) which exit the Hubble radius during AD. They will enter the Hubble radius during DB and will again exit the Hubble radius during BC. Let us calculate the total number of modes which cross the Hubble radius in the time interval (t_1, t_2) or, more conveniently, when the expansion factor is in the range (a_1, a_2). Since the phase space density of the number of modes in the *comoving* Hubble volume $V_{com} = 4\pi/3H^3a^3$ is given by the integral of $dN = V_{com}\, d^3k/(2\pi)^3 = V_{com}k^3/(2\pi^2)\, d\ln k$ we need to compute the integral over the relevant range of k. (We, of course, get the same expression if we use the proper volume and proper wave number instead of comoving variables). If we take the condition for horizon crossing to be $k = Ha$, then we obtain

$$\begin{aligned} N(a_1, a_2) &\equiv \int \frac{d^3x d^3k}{(2\pi)^3} = \int \frac{V_{com}\, d^3k}{(2\pi)^3} \\ &= \frac{2}{3\pi} \ln\left(\frac{H_2 a_2}{H_1 a_1}\right) , \end{aligned} \tag{12.74}$$

where we have used $V_{com} = 4\pi/3H^3a^3$ and $k = Ha$. Note that, as we have defined it, $N(a_1, a_2)$ is transitive with $N(a_1, a_3) = N(a_1, a_2) + N(a_2, a_3)$ with the sign of $N(a_1, a_2)$ being positive if $H_2a_2 > H_1a_1$ and negative otherwise. Further, if a particular mode exits the Hubble radius at $a = a_{exit}$ and enters again at $a = a_{enter}$, then $N(a_{exit}, a_{enter}) = 0$. In fact, for any such mode, $N(a_{exit}, a_F) = -N(a_F, a_{enter})$. Hereafter, it is convenient to choose a_1 and a_2 in $N(a_1, a_2)$ such that $a_1 < a_2$ or $a_1 > a_2$, depending on whether the mode is exiting the Hubble radius or entering the Hubble radius so as to make $N(a_1, a_2)$ positive. (Note that during exit, $k_1^{-1} > k_2^{-1}$ implies $a_1 < a_2$, while during the entry $k_1^{-1} > k_2^{-1}$ implies $a_1 > a_2$.) Then it is clear that during each of the three phases of the universe shown in Fig. 12.2, the total number of modes which cross the Hubble radius remains constant

$$N(a_I, a_F) = N(a_\Lambda, a_F) = N(a_\Lambda, a_{vac}) \equiv N , \tag{12.75}$$

which can be thought of as some kind of conservation law. It is easy to find its value in any of the phases. In the de Sitter phase with constant H we have $Ha \propto a$, while in the radiation dominated phase $H \propto a^{-2}$, so that $Ha \propto 1/a$. Therefore,

$$N = \frac{2}{3\pi} \ln Q \,. \tag{12.76}$$

So the equality of ratios of expansion factors in the three phases translates to the equality of the number of modes N in a Hubble volume which crosses the Hubble radius in the three phases. *This number N is a characteristic, conserved quantity for the universe during the three phases.*

The above relation was obtained by assuming that there is an abrupt change of slope of the Hubble radius at D and B and ignoring the matter domination phase. Correcting for matter domination and including the smooth transition at B is algebraically trivial since we know the behavior of our universe around B. The transition at the end stage of a Planck scale inflation with re-heating, emergence of matter, etc., at D is more uncertain. We have also assumed that the condition for Hubble radius crossing is $k = Ha$. This is equivalent to taking the comoving length scale corresponding to k to be $1/k$; one could have taken this to be $2\pi/k.\pi/k, \ldots$, etc., with the same level of uncertainty. All these effects could introduce order unity corrections to the expression for N and to remind ourselves of this fact we will rewrite the expression in (12.76) as

$$\mu N = \frac{2}{3\pi} \ln Q \,, \tag{12.77}$$

where μ is expected to be a number of order unity. Substituting in (12.73) we can relate the value of the cosmological constant to N by

$$\Lambda L_P^2 = 3Q^{-4} = 3\exp(-6\pi\mu N) \,. \tag{12.78}$$

We consider this an important result in its own right. It reduces the problem of understanding the numerical value of cosmological constant to a more manageable problem of understanding a particular value for μN for our universe.

The natural value for N is just 4π in our scenario. To see this note that during the Planck scale inflation, the surface area of the Hubble sphere is $4\pi L_P^2$ and it is reasonable to assume that the total number of modes crossing this Hubble radius during the Planck scale domain should be of the order of $N_{\text{sur}} = 4\pi L_P^2/L_P^2 = 4\pi$. (There could again be an order unity factor which we will absorb into μ.) Using $N = 4\pi$ in (12.78), we find

$$\Lambda L_P^2 = 3\exp(-24\pi^2\mu) \,. \tag{12.79}$$

We can see that, even for $\mu = 1$, this gives $\Lambda L_P^2 = 3 \times 10^{-103}$, which is within striking distance of the observations and far better than what any other model has achieved. The smallness of the cosmological constant is now related to its exponential dependence on N, plus the fact that $24\pi^2$ is a rather large number! One can obtain the correct, observed value of the cosmological constant for $\mu = 1.18$, which, as advertised, is an order unity number

$$\Lambda_{\text{obs}} L_P^2 = 3\exp(-24\pi^2\mu) \quad (\mu = 1.18) \,. \tag{12.80}$$

We will now make several comments about the results from a broader context.

Usually, one does not consider a Planck scale inflationary scenario because of the claims in the literature that it produces too many gravitational wave perturbations. What is actually provable is that, if one considers spin-2 perturbations *within the framework of normal continuum field theory* in an inflationary background, then the primordial gravitational waves generated will violate the observational bound if the inflation scale is close to Planck scale. However, this is not a convincing argument because, as we go close to the Planck scale, we cannot trust *continuum field theory* of spin-2 field and the results based on it. In fact, there are suggestions in the literature [12.75, 76] that this problem goes away if we considers corrections to propagators arising from quantum gravitational effects in the form of a cutoff at Planck scale. While these are just toy models, it prevents us from taking the gravitational wave bounds seriously to exclude Planck scale inflation.

Second, the computation of N in (12.74) only used one (e.g., scalar) degree of freedom, and one might think that we should multiply it by the effective number of species, g_{eff}, at Planck scale. We do not know what this number is but it turns out to be irrelevant in the picture we have in mind. We consider the transition at D to be the emergence of space along with the emergence of matter degrees of freedom (which leads to the radiation dominated era) from some other pregeometric degrees of freedom. We then expect the equipartition of gravitational and matter degrees of freedom to set the *total* matter degrees of freedom $g_{\text{eff}}N$ to some specific value like 4π. So, the factor g_{eff} does not play any role in the final expression for ΛL_P^2. (That is, it will modify

the intermediate equations (12.76), (12.77), and (12.78) by changing N to $g_{\text{eff}}N$ but the final result in (12.79) will not change when we set $g_{\text{eff}}N = 4\pi$.) In fact, the initial de Sitter phase, with energy scale equal to Planck energy, needs quantum gravitational inputs for its proper treatment and cannot be considered as the usual inflation driven by a scalar field, etc.

Third, there is a curious relation between N and the entropy one can associate with modes that cross the Hubble radius. If $\mathrm{d}N$ is the number of modes which cross the Hubble radius during the period when the expansion factor changes by $\mathrm{d}a$, then they contain the energy $\mathrm{d}E = (k/a)\,\mathrm{d}N$ if we treat the modes as (effectively) massless. If we associate a temperature $T = H/2\pi$ with the Hubble radius, then we can associate an entropy $\mathrm{d}S = \mathrm{d}E/T = 2\pi(k/Ha)\,\mathrm{d}N = 2\pi\,\mathrm{d}N$ with these modes. So the number of modes $N(a_1, a_2)$ which crosses the Hubble radius during $a_1 < a < a_2$ is related to an entropy $S(a_1, a_2) = 2\pi N(a_1, a_2)$. In terms of S, (12.76) becomes $Q = \exp(3\pi N/2) = \exp(3S/4)$. Equivalently

$$e^S = Q^{\frac{4}{3}} = \left(\frac{a_2}{a_1}\right)^{\frac{4}{3}}, \tag{12.81}$$

which relates the expansion factor to this entropy. Also note that, since $aH = \dot{a}$, $S(a_1, a_2)$ relates the expansion rates of the universe at any two epochs directly

$$\begin{aligned}\dot{a}(t_2) &= \dot{a}(t_1)\exp\left[\frac{3\pi}{2}N(a_1, a_2)\right]\\ &= \dot{a}(t_1)\exp\left[\frac{3}{4}S(a_1, a_2)\right].\end{aligned} \tag{12.82}$$

Let us now turn to the conceptual aspects of this approach. I consider the above analysis as a *program capable of determining the numerical value of ΛL_P^2*. Such a program has the following ingredients:

- The universe is described by two fundamental length scales L_P and $\Lambda^{-1/2}$ or – equivalently – by one length scale L_p and the dimensionless ratio ΛL_P^2. This ratio needs to be determined by a physical principle and the fact that it is very small should become obvious when this principle is implemented properly.
- Time translation invariance of the geometry suggests that de Sitter spacetime qualifies as some kind of *equilibrium* configuration. Given the two length scales, one can envisage two de Sitter phases for the universe, one governed by $H = L_P^{-1}$ and the other governed by $H = (\Lambda/3)^{1/2}$. Of these, I would expect the Planck scale inflationary phase to be an unstable equilibrium causing the universe to make a transition towards the second de Sitter phase governed by the cosmological constant. The transient stage is populated by matter emerging along with classical geometry around the point D in Fig. 12.2.
- Such a cosmological model is characterized by a number N related to ΛL_P^2 by (12.78). This N has a direct physical interpretation as the number of modes within a Hubble volume which crosses the Hubble radius during any of the three phases of evolution of the universe. *Because N has a direct physical meaning*, this translates the problem of determining a very small number ΛL_P^2 to the problem of determining a more manageable number $(1/6\pi)\ln(\Lambda L_P^2/3)$, which is of order 10 for our universe.
- We have given an argument as to why N is of order 4π. This has to be a postulate at this stage since we do not understand how matter and space emerge from some pregeometric variables. However, even without such a detailed knowledge one can argue that the numerical value cannot be widely far off from the result $N = 4\pi$. Given a better model for quantum gravity, one should be able to calculate μ and obtain a more precise numerical value for Λ. Incidentally, cosmological observation can be used to determine N and for a wide range of accepted parameter values, we obtain $N \approx 4\pi$.

The acceptance of the arguments in this section provides a route for resolving what is often considered to be a major challenge in theoretical physics. the author believes the final solution to cosmological problem will *only* require refining the various ingredients described in the itemized list above. In particular, we need to accept the existence of two length scales in our universe and look for a first principle argument to determine the ratio between these two scales. For further developments in this direction see [12.77].

12.8 Conclusions

The authors believes that the features of gravitational theories described above makes a strong case for treating gravitational field equations as emergent and having the same conceptual status as equations of fluid dynamics or elasticity. The peculiar features of gravitational field theories all point to such an interpretation and it is fascinating that one could make so much progress without specifying the dynamics of the microscopic degrees of freedom. Further, this perspective offers a refreshingly different paradigm for cosmology and holds hope for determining the numerical value of the cosmological constant, which is considered one of the deepest puzzles in theoretical physics.

References

12.1 T. Padmanabhan: Rep. Prog. Phys. **73**, 046901 (2010), arXiv:0911.5004
12.2 T. Padmanabhan: Lessons from classical gravity about the quantum structure of spacetime, J. Phys. Conf. Ser. **306**, 012001 (2011), arXiv:1012.4476
12.3 T. Padmanabhan: Res. Astron. Astrophys. **12**, 891–916 (2012), arXiv:1207.0505
12.4 P.C.W. Davies: J. Phys. A **8**, 609–616 (1975)
12.5 W.G. Unruh: Phys. Rev. D **14**, 870 (1976)
12.6 A.D. Sakharov: Sov. Phys. Dokl. **12**, 1040 (1968)
12.7 G.E. Volovik: *The Universe in a Helium Droplet* (Oxford Univ. Press, Oxford 2003)
12.8 Hu B. L.: Int. J. Mod. Phys. D, 20, 697 (2011) [arXiv:1010.5837]
12.9 T. Jacobson: Phys. Rev. Lett. **75**, 1260 (1995)
12.10 T. Padmanabhan: Phys. Rev. D **84**, 124041 (2011), arXiv:1109.3846
12.11 C. Lanczos: Z. Phys. **73**, 147 (1932)
12.12 D. Lovelock: J. Math. Phys. **12**, 498 (1971)
12.13 G.W. Gibbons, S.W. Hawking: Phys. Rev. D **5**, 2738 (1977)
12.14 J.W. York Jr.: Boundary terms in the action principles of general relativity. In: *Between Quantum and Cosmos*, ed. by W.H. Zurek, A. van der Merwe, W.A. Miller (Princeton Univ. Press, Princeton 1988) p. 246
12.15 T. Padmanabhan: *Gravitation: Foundations and Frontiers* (Cambridge Univ. Press, Cambridge 2010)
12.16 J.M. Charap, J.E. Nelson: J. Phys. A: Math. Gen. **16**, 1661 (1983)
12.17 O. Miskovic, R. Olea: J. High Energy Phys. **0710**, 028 (2007)
12.18 A. Mukhopadhyay, T. Padmanabhan: Phys. Rev. D **74**, 124023 (2006), hep-th/0608120
12.19 T. Padmanabhan: Dark energy: Mystery of the millennium. Albert Einstein Century Int. Conf. Paris, AIP Conf. Proc. **861**, 858 (2005), astro-ph/0603114
12.20 T. Padmanabhan: Int. J. Mod. Phys. D **17**, 367–398 (2008), gr-qc/0409089
12.21 J.D. Bekenstein: Nuovo Cim. Lett. **4**, 737–740 (1972)
12.22 S.W. Hawking: Commun. Math. Phys. **43**, 199–220 (1975)
12.23 T. Padmanabhan: Phys. Rev. Lett. **78**, 1854 (1997), hep-th-9608182
12.24 T. Padmanabhan: Phys. Rev. **D57**, 6206 (1998)
12.25 T. Padmanabhan: Class. Quantum Gravity **4**, L107 (1987)
12.26 T. Padmanabhan: Ann. Phys. **165**, 38–58 (1985)
12.27 K. Srinivasan, L. Sriramkumar, T. Padmanabhan: Phys. Rev. **D 58**, 044009 (1998), gr-qc-9710104
12.28 T. Padmanabhan: Phys. Rev. D **82**, 124025 (2010), arXiv:1007.5066
12.29 R.M. Wald: Phys. Rev. D **48**, 3427 (1993), gr-qc/9307038
12.30 V. Iyer, R.M. Wald: Phys. Rev. D **52**, 4430 (1995), gr-qc/9503052
12.31 S. Carlip, C. Teitelboim: Class. Quantum Gravity **12**, 1699 (1995)
12.32 S. Massar, R. Parentani: Nucl.Phys. **B575**, 333 (2000)
12.33 A.J.M. Medved, D. Martin, M. Visser: Class. Quantum Grav. **21**, 3111 (2004)
12.34 A.J.M. Medved, D. Martin, M. Visser: Phys. Rev. D **70**, 024009 (2004)
12.35 D. Kothawala, T. Padmanabhan: Phys. Rev. D **79**, 104020 (2009), arXiv:0904.0215
12.36 T. Padmanabhan: Gen.Rel.Grav. **44**, 2681 (2012), arXiv:1205.5683
12.37 T. Padmanabhan: Phys. Rev. D **83**, 044048 (2011), arXiv:1012.0119
12.38 S. Kolekar, T. Padmanabhan: Phys. Rev. D **85**, 024004 (2012), arXiv:1109.5353
12.39 B. R. Majhi, T. Padmanabhan: Noether current from the surface term of gravitational action, Virasoro algebra and horizon entropy, Phys. Rev. D (in press) arXiv:1204.1422
12.40 S. Kolekar, T. Padmanabhan: Phys. Rev. D **82**, 024036 (2010), arXiv:1005.0619
12.41 T. Padmanabhan: Class. Quantum Gravity **19**, 5387 (2002), gr-qc/0204019
12.42 S. Kolekar, D. Kothawala, T. Padmanabhan: Phys. Rev. D **85**, 064031 (2012), arXiv:1111.0973
12.43 T. Padmanabhan: Phys. Rep. **406**, 49 (2005), gr-qc/0311036

12.44 T. Padmanabhan: Gen. Relativ. Gravity **38**, 1547–1552 (2006)
12.45 T. Padmanabhan: Int. J. Mod. Phys. D **15**, 2029 (2006), gr-qc/0609012
12.46 D. Kothawala, S. Sarkar, T. Padmanabhan: Phys. Lett. B **652**, 338 (2007), gr-qc/0701002
12.47 A. Paranjape, S. Sarkar, T. Padmanabhan: Phys. Rev. D **74**, 104015 (2006), hep-th/0607240
12.48 R.G. Cai, S.P. Kim: J. High Energy Phys. **0502**, 050 (2005), hep-th/0501055
12.49 R.G. Cai, L.M. Cao, Y.P. Hu: J. High Energy Phys. **0808**, 090 (2008), arXiv:0807.1232
12.50 R.G. Cai, L.M. Cao, Y.P. Hu, S.P. Kim: Phys. Rev. **78**, 124012 (2008)
12.51 R.G. Cai, L.M. Cao, Y.P. Hu: Class. Quantum Gravity **26**, 155018 (2009), arXiv:0809.1554
12.52 R.G. Cai, L.M. Cao: Phys. Rev. D **75**, 064008 (2007), gr-qc/0611071
12.53 M. Akbar, R.G. Cai: Phys. Rev. D **75**, 084003 (2007), hep-th/0609128
12.54 Y. Gong, A. Wang: Phys. Rev. Lett. **99**, 211301 (2007), arXiv:0704.0793
12.55 D. Kothawala: Phys. Rev. D **83**, 024026 (2011), arXiv:1010.2207
12.56 T. Damour: Quelques proprietes mecaniques, electromagnetiques, thermodynamiques et quantiques des trous noirs (Université Paris, Paris 1979), available online at http://www.ihes.fr/damour/Articles/
12.57 R.H. Price, K.S. Thorne: Phys. Rev. D **33**, 915 (1986)
12.58 T. Padmanabhan: Int. J. Mod. Phys. D **18**, 2189 (2009), arXiv:0903.1254
12.59 T. Padmanabhan: Class. Quantum Gravity **21**, 4485 (2004), gr-qc/0308070
12.60 T. Padmanabhan: Mod. Phys. Lett. A **25**, 1129 (2010), arXiv:0912.3165
12.61 T. Padmanabhan: Phys. Rev. D **81**, 124040 (2010), arXiv:1003.5665
12.62 T. Padmanabhan: Gen. Rel. Grav. **46**, 1673 (2014), arXiv:1312.3253
12.63 T. Padmanabhan: Int. J. Mod. Phys. D **13**, 2293–2298 (2004), gr-qc/0408051
12.64 T. Padmanabhan: Gen. Relativ. Gravity **40**, 2031–2036 (2008)
12.65 T. Padmanabhan, A. Paranjape: Phys. Rev. D **75**, 064004 (2007), gr-qc/0701003
12.66 T. Padmanabhan: Gen. Relativ. Gravity **40**, 529–564 (2008), arXiv:0705.2533
12.67 D. Kothawala, T. Padmanabhan, S. Sarkar: Phys. Rev. D **78**, 104018 (2008), arXiv:0807.1481
12.68 T. Padmanabhan: Adv. Sci. Lett. **2**, 174 (2009), arXiv:0807.2356
12.69 T. Padmanabhan: Class. Quantum Gravity **22**, L107–L110 (2005), hep-th/0406060
12.70 Padmanabhan T.: Emergence and expansion of cosmic space as due to the quest for holographic equipartition (2012) arXiv:1206.4916
12.71 J. Martin: arXiv:1205.3365
12.72 S. Nobbenhuis: arXiv:gr-qc/0411093
12.73 J. D. Bjorken: The Classification of Universes (2004) arXiv:astro-ph/0404233
12.74 D. Lohia: J. Phys. A A **11**, 1335 (1978)
12.75 T. Padmanabhan: Phys. Rev. Lett. **60**, 2229 (1988)
12.76 T. Padmanabhan, T.R. Seshadri, T.P. Singh: Phys. Rev. D **39**, 2100 (1989)
12.77 H. Padmanabhan, T. Padmanabhan: Int. J. Mod. Phys. D **22**, 1342001 (2013)

13. Spacetime and the Passage of Time

George F. R. Ellis, Rituparno Goswami

This paper examines the various arguments that have been put forward suggesting either that time does not exist, or that it exists but its flow is not real. We argue that:

1. Time both exists and flows.
2. An evolving block universe (*EBU*) model of spacetime adequately captures this feature, emphasizing the key differences between the past, present, and future.
3. The associated surfaces of constant time are uniquely geometrically and physically determined in any realistic spacetime model based in general relativity theory.
4. Such a model is needed in order to capture the essential aspects of what is happening in circumstances where initial data does not uniquely determine the evolution of spacetime structure because quantum uncertainty plays a key role in that development.

Assuming that the functioning of the mind is based in the physical brain, evidence from the way that the mind apprehends the flow of time prefers this evolving time model over those where there is no flow of time.

13.1 Spacetime and the Block Universe

In this section we briefly summarize the usual representation in relativity theory of spacetime as an unchanging block universe, and the associated view that the change of time is an illusion.

The nature of spacetime in both special and general relativity has led some people to a view that the passage of time is an illusion [13.1–4]. Given data at an arbitrary time, it is claimed that everything occurring at any later or earlier time can be uniquely determined from that data, evolved according to deterministic local physical laws (this is formalized in standard existence and uniqueness theorems [13.5]). Consequently, there

cannot be anything special about any particular moment; there is no special *now* that can be called the present. Past, present, and future are equal to each other, for there is no surface that can uniquely be called the present.

Such a view can be formalized in the idea of a *block universe* [13.1, 6, 7]: space and time are represented as merged into an unchanging spacetime entity, with no particular space sections identified as the present and no evolution of spacetime taking place. The universe just *is*; a fixed spacetime block, representing all events that have happened and that will happen. This representation implicitly embodies the idea that time is an illusion; time does not *roll on* in this picture. All past and future times are equally present, and the present *now* is just one of an infinite number. *Price* [13.8] and *Barbour* [13.9] in particular advocate such a position. Underlying this, as emphasized by Barbour, is the idea that time-reversible Hamiltonian dynamics provides the foundation for physical theory in general and gravitation in particular. Occasionally cosmology or astrophysics takes into account time-irreversible physics, for example nucleosynthesis in the early universe or the late phases of gravitational collapse, but the notion of the present as a special time remains absent.

The problem with this view is that it is a profound contradiction to our experiences in everyday life, and in particular to the way science is carried out. Scientific theories are developed and then tested by an ongoing process that rolls out in time: initially the theory does not exist; it is developed, tested, refined, finally perhaps accepted. Is it really plausible that all of this process is an illusion, as some claim? Can it really be that *time is real but flow is not* (*Davies* [13.1]), or *time does not exist* (*Barbour* [13.9], *Rovelli* [13.10])? If time is an illusion, how can the mind generate this illusion, when (assuming the validity of present day neuroscience) the mind is based in the brain – a physical entity, governed by the laws of physics?

By contrast to this view, *Broad* already in 1923 [13.11] argued that the true nature of spacetime is best represented as an *emergent block universe* (EBU), a spacetime which grows and incorporates ever more events, *concretizing* as time evolves along each world line [13.12]. Unlike the standard block universe, it adequately represents the differences between the past, present, and future, and depicts the change from the potentialities of the future to the determinate nature of the past. This is the view we present in this chapter – the claim *time is an illusion* results from using an inadequate model of spacetime.

13.2 Time and the Emerging Block Universe

In this section we summarize the alternative representation of spacetime as an ever-growing emergent block universe, embodying the view that the ongoing flow of time is a key physical aspect of reality, and relate this to the classical physics concept of the nature of time.

How do we envisage spacetime and the objects in it as time unrolls? To motivate the EBU model of reality, consider the following scenario [13.12]: a massive object has rocket engines attached at each end that allow it to move either left or right. The engines are fired alternately by a computer, which produces firing intervals and burn times based on a sensor activated by the random decays of a radioactive element [13.13]. These signals select the actual spacetime path of the object from the set of all possible paths. Due to the quantum uncertainty inherent in radioactive decay [13.14–16], the path realized is not determined by the initial data at any previous time; which potential path becomes actual cannot be predicted, it is determined as it happens (Fig. 13.1).

Because the objects are massive and hence produce spacetime curvature, spacetime structure itself is undetermined until the object's motion is determined by the specific radioactive decay that takes place. Instant by instant, the spacetime structure changes from indeterminate to definite. Thus a definite spacetime structure comes into being as time evolves. The random element introduced through the irreducible uncertainty of quantum events ensures that there is no way the future spacetime can be predicted from the past; what will actually happen is not determined until it happens. Second by second, one specific evolutionary history out of all possibilities is chosen, takes place, and becomes cast in stone (sometimes literally). The future is uncertain and indeterminate until local determinations have taken place at the spacetime event *here and now,* designating the present on a world line at a specific instant; thereafter this event is in the past, having become fixed and immutable, with a continually new event on the world line designating the present.

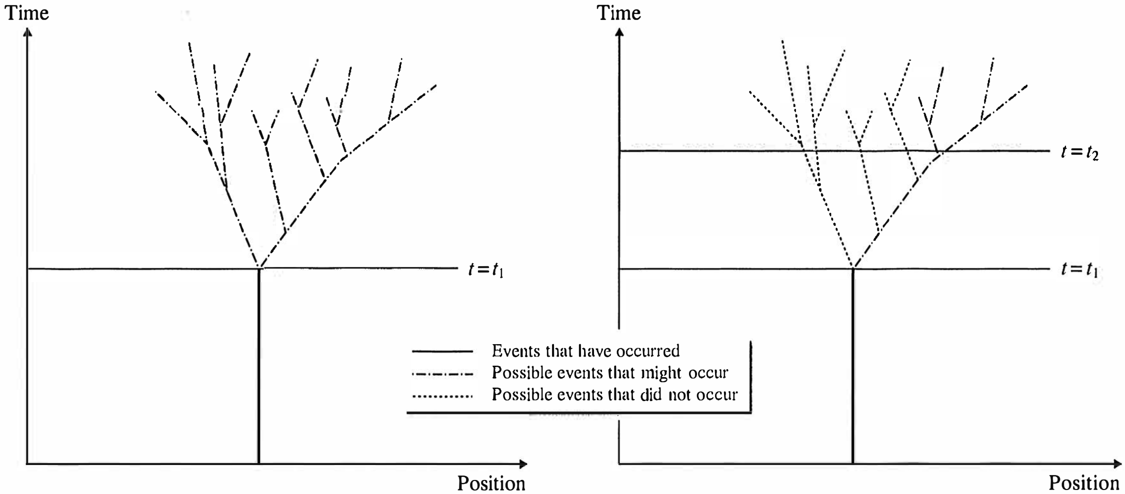


Fig. 13.1 Motion of a particle world line controlled in a random way, so that what happens is determined as it happens. *On the left* events are determined until time t_1 but not thereafter; *on the right*, events are determined until time $t_2 > t_1$, but not thereafter. Spacetime is unknown and unpredictable even in principle before it is determined (because that choice is based on the randomness of quantum decay of radioactive particles). The time at which it is determined inexorably moves on, and given this physical context, this unfolding cannot be stopped, changed, or reversed

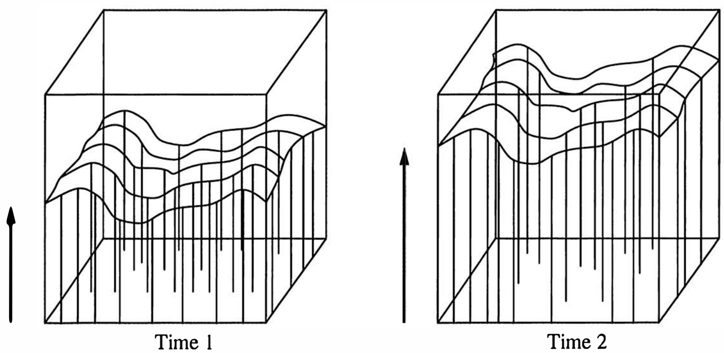


Fig. 13.2 An evolving curved spacetime picture that takes the flow of time seriously. Time evolves along each world line, extending the determinate spacetime as it does so. One cannot locally predict uniquely to the future from data on any constant *time* surface because of quantum uncertainty. This is true both for physics and for the spacetime itself; the developing nature of spacetime is determined by the evolution of the matter in it. A key example is the process whereby quantum fluctuations determine seed spacetime inhomogeneities during the inflationary era in the early universe

The EBU model of spacetime represents this situation (Fig. 13.2): time progresses, events take place, and history is shaped. This is represented through a growing spacetime diagram, in which the past is represented as a usual block universe, but now existing only from the start of spacetime up to the ever-changing surface representing the present. Even the nature of future spacetime, along with the physical events that occur in it, is uncertain; unlike the past, the future does not yet exist, it is just a potentiality; hence it is not represented in the diagram as part of the presently existing spacetime. The passing of time marks the change from indefinite (not yet existing) to definite (having come into being); the present marks the instant at which we can act and change reality. Spacetime grows as time inexorably evolves: at each new in-

stant every previous present has become part of the past [13.12].

The proposed view is thus that spacetime is continually extending to the future as events develop along each world line in a way determined by the complex of causal interactions; these shape the future, including the very structure of spacetime itself. The EBU continues evolving along every world line until it reaches its final state, resulting in an unchanging final block universe at the end of time. One might say that then time has changed into eternity. It is this final block universe that is usually represented in spacetime diagrams, but it only exists when time has run its course everywhere.

13.2.1 The Paradox

This model of spacetime is obviously far more in accord with our daily experience than the standard block universe picture; indeed everyday data, including the apparent passage of time involved in carrying out every single physics experiment, would seem to decisively choose the EBU over the block universe. The evidence seems abundantly clear. Why, then, do some physicists prefer the latter? If the scientific method is to abandon a theory when the evidence is against it, why do some hold to it?

This counter viewpoint was put succinctly by *Sean Caroll* in a blog [13.17]:

> *The past and future are equally real. This isn't completely accepted, but it should be. Intuitively we think that the* now *is real, while the past is fixed and in the books, and the future hasn't yet occurred. But physics teaches us something remarkable: every event in the past and future is implicit in the current moment. This is hard to see in our everyday lives, since we're nowhere close to knowing everything about the universe at any moment, nor will we ever be – but the equations don't lie. As Einstein put it,* It appears therefore more natural to think of physical reality as a four dimensional existence, instead of, as hitherto, the evolution of a three dimensional existence.

However, the question is which equations, and when are they applicable? As emphasized so well by *Eddington* [13.18, pp. 246–260], our mathematical equations representing the behavior of macro objects are highly abstracted versions of reality, leaving almost all the complexities out. The case made in [13.19] is that when true complexity is taken into account, the unitary equations leading to the view that time is an illusion are generically not applicable except to isolated micro components of the whole. The viewpoint expressed by Carroll supposes a determinism of the future that is not realized in practice: *inter alia*, he is denying the existence of quantum uncertainty in the universe we experience. However, physics experiments show uncertainty to be a well-established aspect of the universe [13.14, 16], and it can have macroscopic consequences in the real world, as is demonstrated by the historic process of structure formation resulting from quantum fluctuations during the inflationary era [13.20]. These inhomogeneities were not determined until the relevant quantum fluctuations had occurred and then become crystalized in classical fluctuations; and they were unpredictable, even in principle.

Actually, the EBU proposal does not contradict the first part of the Einstein quote given in Carroll's blog. The core issue not touched on in that quote is where the future boundary of the four-dimensional spacetime advocated by Einstein lies. In the usual block universe picture, it is taken to be at the end of time. In the EBU, it corresponds to the ever-changing present time.

The prime issue arising is that the spacetime view of special relativity denies the existence of any preferred time slices, whereas the claimed existence of the present in the EBU is certainly a preferred time surface (at each instant, it is the future boundary of the four-dimensional spacetime). We will deal with this objection in the following section, after first looking at common physics views of the passage of time in the rest of this section. An array of further arguments for the claim *time is an illusion* have been made by philosophers and physicists; these are conveniently summarized in the Spring, 2012 special issue of *Scientific American* [13.1] (See also [13.21–23] for recent reviews of these issues, with references.). We will turn to these in Sect. 13.4. We then consider the way the block universe view relates to theories of the mind (Sect. 13.6): a key problem for that view. Next, we consider how the EBU picture may be altered when one takes quantum issues into account (Sect. 13.2.3), and point out how it relates to the arrow of time issue (Sect. 13.8.1) and solves the chronology protection question (Sect. 13.8.2). Finally, we reflect on the nature of time in relation to the EBU proposal (Sect. 13.9).

13.2.2 The Classical Physics of the Passage of Time

There are no problems with the existence or passage of time in standard physics textbooks on classical me-

chanics, see, for example, *The Feynman Lectures in Physics* [13.24].

Reversible Dynamics

A standard example is a simple harmonic oscillator (*SHO*) with equation of motion

$$F = m\frac{d^2q}{dt^2} = -kq \,, \tag{13.1}$$

and solution

$$q(t) = A\cos(\omega t - \phi) \,,$$
$$p(t) = m\frac{dq}{dt} = -mA\omega \sin(\omega t - \phi) \,, \tag{13.2}$$

where $\omega := \sqrt{\frac{k}{m}}$. As time evolves, the oscillator oscillates, with its state at time t given by $\{q(t), p(t)\}$. Standard texts discussing the SHO do not question the existence or flow of time.

More generally for dynamical systems with N variables x_i

$$dx_i/\,dt = f(x_j) \quad (i,j = 1 - N) \,, \tag{13.3}$$

the solution $x_i(t)$ represents how the variables change as time flows steadily on. (The SHO is of this form if $x_i = x$, $x_2 = p$.) In these cases knowledge of the state at any time t_0 enables deduction of the state at all earlier and later times; the system is time-reversible and predictable, and evolves with time according to this equation (indeed, the very purpose of these equations is to predict this time evolution). A particular case is Hamiltonian dynamics where

$$\frac{dp_i}{dt} = -\frac{\partial}{\partial q_i}\mathcal{H}(q_j(t), p_j(t)) \,, \tag{13.4}$$
$$\frac{dq_i}{dt} = +\frac{\partial}{\partial p_i}\mathcal{H}(q_j(t), p_j(t)) \,, \tag{13.5}$$

$(i,j = 1 - N)$. There may be constraints, perhaps involving first and second spatial derivatives

$$C_m(p_i, q_j, q_{j,k}, q_{j,kl}) = 0 \tag{13.6}$$

$(m = 1 - M)$, where $q_{j,k} := \partial q_j/\partial\, dx^k$. Then these must be preserved under the time evolution

$$\{(13.4), (13.5)\} \quad \Rightarrow \quad \frac{dC_m}{dt} = 0 \,. \tag{13.7}$$

Then provided the constraints are initially satisfied at time t_0, the past and the future are uniquely determined for some time interval $[T_-, T_+]$ containing t_0

$$\{(p_i(t_0), q_i(t_0)) : C_m(t_0) = 0\}$$
$$\Rightarrow \{(p_i(t), q_i(t)) : T_- < t < T_+\} \,. \tag{13.8}$$

The time development of the system is given by these equations. Three comments are made in the following.

Explicit Time Dependence. The case $\mathcal{H}(q_j(t), p_j(t), t)$, where $\partial\mathcal{H}/\partial t \neq 0$ breaks time translation invariance and explicitly invokes preferred times in the dynamics. We exclude this case in what follows.

Limited Prediction Times. Generically one or both of T_-, T_+ will be finite [13.5]. Except for comments on the chronology protection question (Sect. 13.8.2), we will not consider such global issues here.

First Integrals. For any function $f(q_i, p_i, t)$ (13.4) and (13.5) imply the time derivative

$$\frac{df(q_i, p_i, t)}{dt} = \left(\frac{\partial f}{\partial q_i}\right)\left(\frac{dq_i}{dt}\right) + \left(\frac{\partial f}{\partial p_i}\right)\left(\frac{dp_i}{dt}\right) + \frac{\partial f}{\partial\, dt} \tag{13.9}$$
$$= \left(\frac{\partial f}{\partial q_i}\frac{\partial\mathcal{H}}{\partial p_i} - \frac{\partial\mathcal{H}}{\partial q_i}\frac{\partial f}{\partial p_i}\right) + \partial f/\partial\, dt \,. \tag{13.10}$$

Applying this to the Hamiltonian $\mathcal{H}$ itself,

$$d\mathcal{H}(q_i, p_i)/\,dt = \left(\frac{\partial\mathcal{H}}{\partial q_i}\frac{\partial\mathcal{H}}{\partial p_i} - \frac{\partial\mathcal{H}}{\partial q_i}\frac{\partial\mathcal{H}}{\partial p_i}\right) = 0 \,, \tag{13.11}$$

so $\mathcal{H}$ is the conserved energy (related to time translation invariance of the dynamics); in simple cases with kinetic energy $T(p) := \frac{1}{2}p^2$ and potential energy $V(q)$,

$$\mathcal{H}(p(t), q(t)) = T(p) + V(q) = const =: E \,. \tag{13.12}$$

This constant relation does not imply that there is no time evolution taking place; it means there is a first integral of that evolution. If there is no conserved energy, the Hamiltonian description {(13.4) and (13.5)} does not apply. This will be the case whenever dissipative processes take place and affect the dynamics at the chosen scale of description; this occurs in a great many cases in both macro and micro physics [13.25].

Irreversible Dynamics

In general, friction effects mean we have an inability to retrodict if we lose information below some level of coarse graining. The simplest example is a block of mass m sliding on a plane, slowing down due to constant limiting friction $F = -\mu R$, where μ is the coefficient of friction and $R = mg$ is the normal reaction, where g is the acceleration due to gravity [13.26]. The motion is a uniform deceleration; if we consider the block's motion from an initial time $t = 0$, it comes to rest at some later time $t_* > 0$. For $t < t_*$ the velocity $v(t)$ and position $x(t)$ of the object are given by

$$v_1(t) = -\mu g t + v_0 ,$$
$$x_1(t) = -\frac{1}{2}\mu g t^2 + v_0 t + x_0 , \tag{13.13}$$

where (v_0, x_0) are the initial data for (v, x) at the time $t = 0$. This expression shows that it comes to rest at $t_* = \mu g / v_0$. For $t > t_*$, the quantities v and x are given by

$$v_2(t) = 0, x_2(t) = X(\text{constant}) , \tag{13.14}$$

where $X := -\frac{1}{2}\mu g t_*^2 + v_0 t_* + x_0$.

The key point now is that from the later data (13.14) at any time $t > t_*$ you cannot determine the initial data (v_0, x_0), nor even the time t_* when the object came to rest, thus you cannot reconstruct the trajectory (13.13) from that later data. You cannot even tell if the block came from the left or the right. The system is no longer time reversible or predictable. The direction of time is uniquely determined by the way it came to a halt (it cannot spontaneously start moving; note that the rest frame implied here is that defined by the table on which the block slides). That coming to rest was an event that took place in time; time does not cease after it comes to rest, it is the block's motion that ceases then. At macro scales, reversibility does not hold, nor is the motion predictable in both directions of time; the dynamics is not Hamiltonian.

A standard response is, if we knew how the surface was heated, we could work out where the block cane from and when it arrived. However, the heat dissipates away and vanishes into thermal fluctuations (at a small enough scale, into quantum fluctuations); that record is soon irretrievably lost. The claim that *every event in the past and future is implicit in the current moment* soon ceases to be true.

13.2.3 Quantum Physics of the Passage of Time

In the last example, our inability to predict is associated with a lack of detailed information. So if we were to fine-grain to the smallest possible scales and collect all the available data, could we then determine uniquely what is going to happen? No, we cannot predict to the future in this way because of foundational quantum uncertainty relations(see e.g., [13.27–29]). We cannot predict precisely when a nucleus will decay or what the velocity of the resultant particles will be, nor can we predict precisely where a photon or electron in a double-slit experiment will end up on the screen. This unpredictability is not the result of a lack of information; it is the very nature of the underlying physics.

There are two kinds of quantum evolution [13.28–30]. First, there is the unitary Schrödinger evolution

$$\mathrm{i}\hbar \frac{\partial |\psi\rangle}{\partial t} = \hat{H}|\psi\rangle , \tag{13.15}$$

where $|\psi(t)\rangle$ is the wave function and $\hat{H}$ the Hamiltonian operator. This equation determines its evolution: time occurs here in the same way as in classical physics, although the meaning of the relevant variable is quite different. Time rolls on and the state vector evolves; therefore probabilities change with time.

Second, there is wave function reduction, associated with both state vector preparation and measurement events. Consider the wave function or state vector $|\psi(x)\rangle$. The basic expansion postulate for quantum mechanics is that before a measurement is made, $|\psi\rangle$ can be written as a linear combination of eigenstates

$$|\psi_1\rangle = \sum_n c_n |u_n(x)\rangle , \tag{13.16}$$

where u_n is an eigenstate of some observable $\hat{A}$ (see e.g. [13.29, pp. 5–7]). Immediately after a measurement is made at a time $t = t^*$, however, the wavefunction is found to be in one of the eigenstates

$$|\psi_2\rangle = c_N u_N(x) , \tag{13.17}$$

for some specific index N. The data for $t < t^*$ do not determine N; they merely determine a probability for the outcome N through the fundamental equation

$$p_N = c_N^2 . \tag{13.18}$$

One can think of this as being due to the probabilistic time-irreversible reduction of the wave function [13.28, pp. 260–263]

$$|\psi_1\rangle = \sum_n c_n |u_n(x)\rangle \longrightarrow |\psi_2\rangle = c_N u_N(x)\,. \quad (13.19)$$

This is the event where the uncertainties of quantum theory become manifest, as the indeterminate future makes a transition to the determined past (up to this time the evolution is determinate and time reversible). Invoking a many-worlds description (see, e.g., [13.29]) will not help determine the specific outcome; in the actually experienced universe in which we make the measurement, N is unpredictable, as confirmed by experiment. The specific experimental outcome (13.17) that will be measured by an observer to occur at a later time is not determined by the Everett hypothesis. (This assumes that in Fig. 13.1, *all* those possible paths in fact occurred; but we experience only one specific path.)

Thus the initial state (13.15) does not uniquely determine the final state (13.17), and this is not due to lack of data, it is due to the foundational nature of quantum interactions. You can predict the statistics of what is likely to happen but not the unique actual physical outcome, which unfolds in an unpredictable way as time progresses; you can only find out what this outcome is after it has happened. Furthermore, in general the time t^* is not predictable from the initial data either; you do not know when the *collapse of the wave function* (the transition from (13.15) to (13.17)) will happen (you cannot predict when a specific excited atom will emit a photon or a radioactive particle will decay).

We also cannot retrodict to the past at the quantum level, because once the wave function has collapsed to an eigenstate we cannot tell from its final state what it was before the measurement. You cannot retrodict uniquely from the state (13.17) immediately after the measurement takes place, or from any later state that it then evolves to via the Schrödinger equation at later times $t > t^*$, because knowledge of these later states does not suffice to determine the initial state (13.15) at times $t < t^*$; the set of quantities c_n are not determined by the single number a_N.

This process takes place all the time as physical events occur and have classical outcomes (in photosynthesis in plants and in nucleosynthesis in the early universe, for example); it is not necessarily associated with a measuring apparatus or the mind of an experimenter. However, it is time-irreversible, causing information loss, and so is not describable by any unitary evolution. The classical world would not exist if this did not happen as an ongoing unfolding process in time.

The fact that such events happen at the quantum level does not prevent them from having macro-level effects. Many systems can act to amplify them to macro levels, including photomultipliers (whose output can be used in computers or electronic control systems). This amplification is what occurred when cosmic rays – whose emission is subject to quantum uncertainty – caused genetic damage in the distant past, resulting in new phenotypes occurring [13.31]. The specific outcome that actually occurred was determined as it happened, when quantum emission of the relevant photons took place. Any specific emission event (a photon emission time and trajectory) was not determined by the priori quantum state, so any consequent damage to a specific gene in a particular cell at a particular time and place cannot be predicted even in principle. (This damage is not trivial, see [13.32].) Consequently the specific evolutionary outcomes for life on Earth (the existence of dinosaurs, giraffes, humans) cannot be uniquely determined by causal evolution from detailed data at the start of life on Earth.

13.3 A Problem: Surfaces of Change

The problem, however, is the claimed unique status of *the present* in the EBU – the surface where the indeterminate future is changed to the definite past at any instant. In this section, we propose that there are indeed such preferred surfaces in all realistic general relativity models of spacetime.

It is a fundamental feature of special relativity that simultaneity is not uniquely defined; it depends on the state of motion of the observer, and one presumes that in some fundamental sense the present must be regarded as a surface of constant time. What is past and future elsewhere depends on one's motion, hence the block universe model is natural; it is the only way a spacetime model can incorporate this lack of well-defined surfaces of instantaneity. For different observers at an event P, different surfaces of simultaneity will designate different events Q on a distant world line γ_1 as simultaneous with P [13.33]; the only resolution is that they are all simultaneous with P, hence time is an illusion.

However, there are two fundamental points to be made here that completely change the picture. First, the physical events that shape how things evolve are based on particle interactions, and take place along time-like or null world lines, not on space-like surfaces, which are secondary. The concept of simultaneity is only physically meaningful for neighboring events; it has no physical impact for distant events, it is merely a theoretical construct we like to make in our minds. What we think is instantaneous makes no difference to our interaction with a vehicle on Mars. What is significant is firstly what happens over there, secondly what happens here on Earth, and, thirdly the signals between us. Simultaneity does not enter into it.

What really matters is proper time τ measured along timelines $x^i(\nu)$, determined from the metric tensor $g_{ij}(x^k)$ by the basic formula [13.5, 33]

$$\tau = \int \sqrt{-\mathrm{d}s^2} = \int \sqrt{-g_{ij}(\mathrm{d}x^i/\mathrm{d}\nu)(\mathrm{d}x^j/\mathrm{d}\nu)}\,\mathrm{d}\nu\,. \tag{13.20}$$

Indeed this is the reason why the metric tensor is central to relativity theory: as well as determining which lines are null lines ($\mathrm{d}\tau = 0$ all along the curve), it determines proper time along time-like world lines. Natural surfaces of constant time have been given by this integral since the start of the universe. Thus we can propose the following.

The Present

The ever-changing surface $S(\tau)$ separating the future and past – the *present* – at the time τ is the surface $\{\tau = constant\}$ determined by the integral (13.20) along a family of fundamental world lines starting at the beginning of spacetime.

If the universe were to exist forever we would have to start at some arbitrarily chosen *present* time $\{\tau_0 = const\}$, which we assume exists, and integrate from there.

However, is this well defined, given that there are no preferred world lines in the flat spacetime of special relativity? The second fundamental feature is that it is general relativity that describes the structure of spacetime, not special relativity. Gravity governs spacetime curvature [13.5], and because there is no perfect vacuum anywhere in the real universe (*inter alia* because cosmic blackbody background radiation permeates the Solar System and all of interstellar and intergalactic space [13.20]), spacetime is nowhere flat or even of constant curvature; therefore there are preferred time-like lines everywhere in any realistic spacetime model [13.34]. The special relativity argument does not apply.

A unique geometrically determined choice for fundamental world lines is the set of time-like eigenlines $x^a(\nu)$ of the Ricci tensor (they will exist and be unique for all realistic matter, because of the energy conditions such matter obeys [13.5]). Their four-velocities $u^a(\nu) = \mathrm{d}x^a(\nu)/\mathrm{d}\nu$ satisfy

$$T_{ab}u^b = \lambda_1 u_a \Leftrightarrow R_{ab}u^b = \lambda_2 u_a\,, \tag{13.21}$$

where the equivalence follows from the Einstein field equations. Thus we can further propose the following.

Fundamental World Lines

The proper time integral (13.20) used to define the present is taken along the world lines with four-velocity $u^a(\nu)$ satisfying (13.21).

In effect this is the proper time comoving gauge used in perturbation theories; it will, of course, give the usual surfaces of constant time in the standard Friedmann–Lemaître–Robertson–Walker (FLRW) cosmologies.

The following two key issues arise regarding this proposal:

- *What about general covariance and local Lorentz invariance?* These are symmetries of the general theory, not of its solutions. Interesting solutions break the symmetries of the theory; this is not surprising, as we know that broken symmetries are the key to interesting physics [13.35].
- *What about simultaneity?* In general these surfaces are not related to simultaneity as determined by radar [13.33]; indeed this is even so in the FLRW spacetimes (where the surfaces of homogeneity are generically not simultaneous, according to the radar definition [13.36]). The flow lines are not necessarily orthogonal to the surfaces of constant time; indeed they may have nonzero vorticity and acceleration as well as shear and expansion, so there may be no surfaces orthogonal to the flow lines [13.34]. More than that, the surfaces determined in this way are not even necessarily space-like, in an inhomogeneous spacetime.

The latter feature means that the there may possibly be a *time horizon*; a null boundary where these surfaces make a transition from space-like to time-like. This will of course only happen for very large gravitational fields such as occur in black holes; indeed, these surfaces may

well usually coincide with an event horizon. The initial value problem will then be very different when based in data in these surfaces; however even if these surfaces become time-like (necessarily then being null in some places), data on them will still determine the spacetime in their future and past Cauchy development, up to intersections of this development with surfaces where the outcome is already determined. The physics of time then will be quite different than usual: this needs investigation. It could relate to the black hole information paradox.

In Summary. While the general coordinate invariance invoked in general relativity theory might be thought to proclaim there are no preferred such surfaces, in any particular solution this is not the case – there will be preferred time-like lines in any realistic cosmological solution. The result will be existence of a family of preferred surfaces representing constant proper time τ since the start of the universe along these fundamental world lines. The proposal is that each represents what was the *present* at the corresponding time τ, for all times up to the present time τ_0 (they do not exist for $\tau > \tau_0$, for that spacetime is not yet determined). These surfaces are derivative rather than primary, as they result from the configuration of fundamental world lines. They will usually not be instantaneous as determined by radar soundings.

13.4 Other Arguments Against an EBU

A series of other arguments, both physical and philosophical, have been deployed in favor of the standard block universe picture [13.1], and hence deny the EBU proposal. In this section we review these and argue that none of them are fatal.

13.4.1 Categorization Problem

A philosophical argument is that the past, present, and future are exclusive categories, so a single event cannot have the character of belonging to all three. The counterargument is as follows:

- Suppose E happens at t_E.
- At time $t_1 < t_E$, E is in future,
- At time $t_1 = t_E$, E is in present,
- At time $t_1 > t_E$, E is in past.

Its category changes – that is the essence of the flow of time – so this is a semantic problem, not a logical one. One needs adequate semantic usage and philosophical categories to allow description of this change: language usage cannot prevent the flow of time.

13.4.2 Not Necessary to Describe Events

Davies [13.1] and *Rovelli* [13.10] claim time does not flow because it is not needed to describe the relations between relevant variables, which are all that matter physically. Thus one can always obtain correlations between position $p(t)$ and momentum $q(t)$ for a system by eliminating the time variable: solve for $t = t(q)$ and then substitute to obtain $p(t) = p(t(q)) = p(q)$, and time has vanished! For example, in the case of the simple harmonic oscillator (13.2), this gives the SHO phase plane

$$q^2 + \left(\frac{p}{m}\right)^2 = A^2 \,. \tag{13.22}$$

Thus one can describe system changes by relating component variables to one another, rather than to a global idea of time, which suggests nothing happens or changes, they are just correlated.

Yes indeed, one can find this time-independent representation of what happens. (This is just the energy integral (13.12).) However, that does not mean that time does not flow, it just means that the results of times flow are correlations between relevant variables. That abstraction represents part of what happens, namely the relation between p and q, and omits other parts, namely the relation to time. One can put time back to obtain

$$q(t)^2 + \left(\frac{p(t)}{m}\right)^2 = A^2 \,, \quad q(t) = A\cos(\omega t - \phi) \,, \tag{13.23}$$

and the point representing the system moves along the flow lines as time changes. The first model leaves out part of what is happening: that does not mean it does not happen, it just means it is a partial model of reality, including some aspects and omitting others. It leaves out the way that the continually changing correlations flow smoothly one after another in a continuous ongoing way.

13.4.3 Rates of Change

A key question is *What determines the rate of flow of time?* or *How fast does time pass?* Davies and others suggest there is no sensible answer to this question. In contrast, we claim that the answer is given by (13.20), which determines proper time τ along any world line. This is the time that will be measured along that world line by any perfect clock [13.5, 33]; real world clocks – oscillators that obey the simple harmonic equation – are approximations to such ideal clocks, and it is the relation between such clocks and other physical events that measures the passage of time.

The Preferred Time Parameter

The whole edifice of physics is built on the assumption that we can build such clocks to a good approximation, giving a time parameter τ that appears equally in all dynamical equations of physics: Newton's equation of motion [13.24], Maxwell's equations [13.37], the Schrödinger equation [13.14], General relativity expressed in a $1+3$ covariant formalism [13.34], and so on; because of this, a lot of money is spent on building idealized clocks that are understood to be more accurate than any previous clock (see, e.g., http://en.wikipedia.org/wiki/Clock; accurate navigation, for example, requires accurate timekeeping [13.38], which is thus a core feature of GPS systems). Standard physics would not work if a different time parameter was needed in each of these equations. Special and general relativity identify that time as proper time (13.20) along time-like worldlines.

Given such clocks, the rate of change with time of any variable $f(\tau)$ along a world line is given by

$$f' = \mathrm{d}f/\mathrm{d}\tau \,. \tag{13.24}$$

Then choosing $f(\tau) = \tau$, the answer to the question posed is that *the rate of change of time is unity*

$$\tau' = \mathrm{d}\tau/\mathrm{d}\tau = 1 \,. \tag{13.25}$$

In other words, through (13.20) the rate of change of time in any particular coordinate system is determined by the metric tensor. Using normalized comoving coordinates with $u^a = \delta^a_0$ and the time parameter v chosen as τ [13.34], $g_{00} = -1$, and (13.20) becomes

$$\tau = \int \mathrm{d}\tau \,. \tag{13.26}$$

The relative flow of time along different world lines may be different; that is the phenomenon of time dilation, caused by the varying gravitational potentials represented by the metric tensor [13.39]. However, this does not mean it is not well defined along each world line.

The Metric Evolution

So if the metric tensor determines proper time, what determines the metric tensor? The Einstein field equations, of course [13.5]. These can be expressed in many ways, for example a 1+3 covariant formalism [13.34], a tetrad formalism [13.40], or ADM (Arnowitt, Deser, and Misner) formalism [13.41]. Following the ADM approach, the first fundamental form (the metric) is represented as

$$\mathrm{d}s^2 = (-N^2 + N_i N^i)\,\mathrm{d}t^2 + N_i\,\mathrm{d}x^i\,\mathrm{d}t + g_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j \,, \tag{13.27}$$

where $i, j = 1, 2, 3$. The lapse function $N(x^\alpha)$ and shift vector $N_i(x^\beta)$ represent coordinate choices, and can be chosen arbitrarily; $g_{ij}(x^\alpha)$ is the metric of the three-spaces $\{t = const\}$. The second fundamental form is

$$\pi_{ij} = n_{i;j} \,, \tag{13.28}$$

where the normal to the surfaces $\{t = const\}$ is $n_i = \delta^0_i$; the matter flow lines have tangent vector $u^i = \delta^i_0$ (which differs from $n^i = g^{ij} n_j$ whenever $N_i \neq 0$, cf. [13.42]). The field equations for g_{ij} are as follows (where three-dimensional quantities have the prefix (3)): four constraint equations

$$^{(3)}R + \pi^2 - \pi_{ij}\pi^{ij} = 16\pi\rho_{\mathrm{H}} \,, \tag{13.29}$$

$$R^\mu := -2\pi^{\mu j}{}_{|j} = 16\pi T^\mu_0 \,, \tag{13.30}$$

where $|j$ represents the covariant derivative in the three-surfaces, and 12 evolution equations

$$\partial_t g_{ij} = 2Ng^{-1/2}\left(\pi ij - \tfrac{1}{2} g_{ij}\pi\right) + N_{i|j} + N_{j|i} \,, \tag{13.31}$$

$$\begin{aligned}\partial_t \pi_{ij} = {} & -Ng^{-1/2}\left({}^{(3)}R_{ij} - \tfrac{1}{2} g_{ij}{}^{(3)}R\right) \\ & + \tfrac{1}{2} Ng^{-1/2} g_{ij}\left(\pi_{mn}\pi^{mn} - \tfrac{1}{2}\pi^2\right) \\ & - 2Ng^{-1/2}\left(\pi^{im}\pi_m{}^j - \tfrac{1}{2}\pi\pi^{ij}\right) \\ & + \sqrt{g}\left(N^{|ij} - g^{ij}N^{|m}{}_{|m}\right) + (\pi^{ij}N^m)_{|m} \\ & - N^i_{|m}\pi^{mj} - N^j_{|m}\pi^{mi} + 16\pi\hat{T}_{ij} \,.\end{aligned} \tag{13.32}$$

Equations of state for matter must be added, and the matter conservation equations $T^{ab}{}_{;b} = 0$ satisfied (as is required for consistency of the evolution equations). Then (13.31) determines the rate of change of the metric $g_{ij}(x^\alpha)$ relative to the ADM time coordinate; (13.32)

determines the rate of change of the geometric source terms π_{ij} occurring in (13.31); and the matter equations determine the rate of change of the matter terms. How this works out in practice is shown in depth in [13.43]. Overall this determines the metric tensor as a function of time, and hence evolution of the surfaces of constant time as defined above (which are determined by the metric).

This can be worked out using *any* time surfaces (that is the merit of the ADM formalism); in particular one can specialize the time surfaces and flow lines to those defined above (Sect. 13.3):

1. We choose the 4-velocity to be a Ricci Eigenvector:
$$T_0^{\mu} = 0 \Rightarrow R^{\mu} = -2\pi^{\mu j}{}_{|j} = 0\,, \qquad (13.33)$$
which algebraically determines the shift vector $N_i(x^j)$, thereby solving the constraint (13.30).
2. We determine the lapse function $N(x^i)$ by the condition that the time parameter t measures proper time τ along the fundamental flow lines.

These conditions uniquely determine the lapse and shift (see the Appendix for details). Then, given the equations of state and dynamical equations for the matter (in the case of a perfect fluid, [13.43, (45)–(77)]) with $S^{\kappa} = S\delta_0^{\kappa}$, (13.31) and (13.32) determine the time evolution of the metric in terms of proper time τ along the fundamental flow lines; the constraints are conserved because of energy-momentum conservation. The development of spacetime with time takes place just as is the case for other physical fields, with the relevant time parameter being proper time along the fundamental flow lines. There is no problem with either the existence or the rate of flow of time.

Predictability

Do these equations mean the spacetime development is uniquely determined to the future and the past from initial data? That all depends on the equations of state of the matter content: the relations between the density ρ_{H} in (13.29), pressure tensor $\hat{T}_{ij}$ in (13.32), and momentum density T_0^{μ} in (13.30). These quantities depend on the frame chosen, and T_0^{μ} is zero when we make the choice (13.33).

Assuming this choice, define the pressure p and anisotropic stress Π_{ij} by

$$p = \frac{1}{3} g^{ij}\hat{T}_{ij}\,, \quad \Pi_{ij} = \hat{T}_{ij} - p g_{ij}\,. \qquad (13.34)$$

The future and past will be uniquely determined for simple equations of state such as noninteracting baryons plus radiation, as in standard cosmological models

$$\rho_{\mathrm{H}} = \rho_b + \rho_r\,, \quad p = p_b + p_r = \tfrac{1}{3}\rho_r\,, \quad \Pi_{ij} = 0\,, \qquad (13.35)$$

where the energy density conservation equations will determine the time evolution of ρ_b, ρ_r. However:

- One can have dissipative processes (shear viscosity or bulk viscosity, for example [13.34]), so the evolution is not time reversible – a Hamiltonian description does not apply to the matter, and hence does not apply to the combined (matter, gravity) system either, where matter determines space-time curvature.
- One can have an explicitly time-dependent equation of state
$$p = p(\rho_{\mathrm{H}}, \tau)\,, \quad \Pi_{ij} = \Pi_{ij}(\rho_{\mathrm{H}}, \tau)\,, \qquad (13.36)$$
and predictability is no longer the case if the time dependence is not predictable from the available initial data at the relevant scales; again a time-reversible and predictive Hamiltonian description cannot be used for the system as a whole.

For example, one can have a massive body where effectively one has

$$\Pi_{ij}(\tau) = F(\tau)\Pi_{ij}(0)\,, \qquad (13.37)$$

where $F(\tau)$ represents internal dynamics not visible to the external world (Bondi's massive objects Tweedledum and Tweedledee incorporated this idea, see *Narlikar* [13.44] for a description); there might be a mechanism here that is computer controlled via an algorithm embodying a random number generator (see [13.45] for a discussion.) or based in random signals generated via radioactive decay [13.13]. Then as explained above (Sect. 13.2) and in Fig. 13.1, the initial data do not determine the outcome ($F(\tau)$ indicates a causal influence but not a predictable functional relation). One can only determine what will happen as it happens.

So (13.31) and (13.32) determine the time evolution of the spacetime, but do not guarantee predictability either of the future or the past. That depends on the physics of the matter; if quantum unpredictability is amplified to macro scales, the spacetime evolution is intrinsically undetermined until it happens (as mentioned above, this was essentially what happened during the generation of seed inhomogeneities in the inflationary era in the very early universe).

Conclusion

Time flows at the rate of one second per second, with the metric tensor determining what this rate is for clocks and every other physical system (the choice of units is, of course, arbitrary, but can be done in a way that makes sense). The result is that clock readings and particle motions are correlated in a way that enable us to reliably predict motions. So yes, such correlations are fundamental to our experience of the flow of time (as emphasized by Davies); they are a result of its equally inexorable omnipresent continuing flow in all physical systems. Can you change the rate of time? No! Can you stop time? No! Can you reverse time? No! Like Old Man River, it just keeps flowing on; that is the primitive expressed in all the time evolution equations of physics [13.14, 24, 34, 37] and the related existence and uniqueness theorems [13.5].

13.4.4 Time Parameter Invariance of General Relativity

What about the time parameter invariance of general relativity, as made manifest in the ADM formalism [13.41, 46]? This has basically already been dealt with in the discussion above:

- The gravitational side of the ADM equations may be time-parameter invariant, but the matter side is not, in particular because rescaling time changes the value of the kinetic energy $T(p)$ while leaving the potential energy $U(x^i)$ unchanged. Hence any solutions with matter present (i. e., all realistic solutions) will not be time parameter invariant; this is part of the ongoing tension between the geometric and matter sides of the Einstein field equations.
- This is part of the broader theme mentioned above: specific solutions of the theory have less symmetry than the theory itself; this symmetry breaking is a key feature of all realistic solutions of the equations of physics [13.35], and in particular cosmological solutions (Sect. 13.3).
- Proper time (13.20) along fundamental world lines (13.21) provides a preferred time parameter in realistic solutions of general relativity theory.

Local physics does indeed have a preferred time parameter: for example, in a simple harmonic oscillator using standard time t, $q(t) = A\cos(\omega t - \phi)$ (see (13.2)); these cycles measure time t like a metronome (which is why SHOs are used as clocks). One can, in principle, change to an arbitrary time t',

$$t' = t'(t) \Rightarrow t = t(t') \Rightarrow q = A\cos(\omega t(t') - \phi)\,, \tag{13.38}$$

for example,

$$\begin{aligned} t' &= \exp H(t) \Rightarrow t \\ &= \exp(-Ht') \Rightarrow q \\ &= A\cos(\omega \exp(-Ht') - \phi)\,, \end{aligned} \tag{13.39}$$

hence the regular motion no longer is represented as regular. This is because one has chosen a peculiar time, which will not correlate simply with any other physical behavior [13.46, p. 57–60]. The sensible choice of time is that which makes sense of patterns of physics behavior; so the maximum sensible variation is $t' = \alpha t + \beta$ (just as in the case of the allowed change in affine parameter along geodesics). One can choose any other reparametrization of time; but any transformations other than affine transformations of proper time confuse and hide what is actually happening. One is able to choose proper time and does so if one wants to illuminate the physics.

It is nonoptimal to examine the dynamics of general relativity without acknowledging the central role of the metric tensor g_{ij} and resultant proper time (13.20) along world lines – which are at the core of the physical interpretation of general relativity [13.5, 33]. You can use a proper time coordinate τ in the ADM formalism (as shown above in Sect. 13.4.1; this is a general way of solving the problem of time in ADM dynamics, see [13.47]). That choice ties this time parameter in to the rest of physics and in particular to time as measured by local ideal clocks (such as a cesium atom). Consequently, the flow of time is then characterized by the relation between such clocks and other physical events, including the gravitational dynamics represented by the ADM evolution equations (13.31), (13.32). In this sense time is relational (cf. [13.46, p. 163–166]).

In the case of the standard FLRW models of cosmology, the usual [13.5] metric is

$$\mathrm{d}s^2 = -\mathrm{d}\tau^2 + a^2(\tau)\,\mathrm{d}\sigma^2\,, \tag{13.40}$$

where $\mathrm{d}\sigma^2(x^i)$ is a time-independent three-space of constant curvature [13.20, 34] and τ is precisely the preferred time coordinate defined above (Sect. 13.3), as the flow lines $u^a = \delta^a_0$ are Ricci flow lines and τ is

proper time along them (these are normalized comoving coordinates [13.34]). Just as in the case of the simple harmonic oscillator, one can choose other, arbitrary time parameters – but it is very perverse to do so, and this is not done in practice, except for one other common choice: the use of a conformal time parameter $\eta(\tau)$

$$\eta = \int \mathrm{d}\tau / a(\tau) \Rightarrow \mathrm{d}s^2 = a^2(\eta)(-\mathrm{d}\eta^2 + \mathrm{d}\sigma^2) \,, \tag{13.41}$$

which is a very poor representation of proper time but represents causal structure very well [13.5, 33]. However, this still has the same surfaces of constant time as in (13.40) and is always acknowledged to be different from the proper time τ that is fundamental in local physics. One never finds proposals for a time $t = t(\tau, x^i)$ with $\partial t / \partial xi \neq 0$, which would have different surfaces of constant time. The usual choice as in (13.40), agreeing with the proposal made here in (Sect. 13.3), ties time τ in to all the rest of physics, astronomy, geology, technology, and biology.

In Conclusion

Standard physics is based on the choice of a preferred time parameter τ along matter world lines. General relativity both allows such a choice and can itself be written in terms of that choice.

13.5 Time with an Underlying Timeless Substratum

There are a number of proposals for an effective time to somehow emerge in the context of a timeless substratum. These include the Mott proposal (Sect. 13.5.1), the Rovelli proposal (Sect. 13.5.2), and proposals based on the the Wheeler–de Witt equation (Sect. 13.5.3). In this section we will comment on them in turn.

As a preliminary, we first remark that there are two common themes cutting across them all:

- If an effective time emerges at the macro scale, then however that happens, it emerges, and the EBU proposal is then good at macro scales, no matter how it relates to a timeless substrate.
- All of these approaches are based on unitary Schrödinger evolution, so none of them effectively tackles the nonunitary evolution associated with both state vector preparation and quantum measurements [13.29]. Hence they omit a key way that a flow of time takes place at the micro level (Sect. 13.2.3). The many worlds view often associated with the Wheeler–de Witt equation proposes to deal with measurements, but not with state vector preparation, which is also nonunitary.

We regard the latter as a particularly significant problem for all such proposals (cf. [13.19]).

13.5.1 Interaction with the Environment

Mott [13.48] and *Briggs* and *Rost* [13.49, 50] suggest that the time-independent Schrödinger equation (*TISE*) is more fundamental than the time-dependent Schrödinger equation (13.15) (*TDSE*); indeed that the latter emerges from the former, the interaction between parts in a timeless whole generates an effective time characterizing interaction between the parts. This is summarized in [13.50] as follows:

> *Following work of Born, in 1931 Mott (3) described the impact of α-particles on atoms by treating both atom and beam quantum-mechanically with the TISE. Then he showed that for a high energy beam he could describe its motion classically resulting in a time-dependent Hamiltonian and TDSE for the atom alone... time is entering only from a classical interacting environment... time enters the quantum Hamiltonian only when some external system is approximated by classical behavior.*

Thus this is an emergence of time by a top-down interaction from the environment.

According to *Briggs* and *Rost* [13.50], one can use the TIDE in the form

$$H\Psi = E\Psi \Leftrightarrow (H_{\mathcal{E}} + H_S + H_{\mathcal{I}})\Psi = E\Psi \,, \tag{13.42}$$

where $\mathcal{E}$ represents the environment, S the system, and $\mathcal{I}$ their interaction. Through a kinetic energy term, the interaction Hamiltonian somewhat mysteriously introduces an effective time into the wave function for the system. If one accepts this proposal, it is a way that an effective time variation is induced at the quantum level due to top-down effects from the environment – a proposal that is in consonance with the broad suggestions

of the effectiveness of top-down effects in quantum physics presented in [13.19]. Once this has occurred, one has an effective EBU situation at the micro as well as the macro level.

13.5.2 Get It by Coarse Graining?

By contrast *Rovelli* [13.10] suggests time can emerge in a bottom-up way from a timeless substrate as a thermodynamic variable. We find it difficult to see how any process of coarse graining a static state can introduce time, but in any case this proposal faces two other problems.

First, it is based in the idea of equilibrium distribution:

> *Whatever the statistical state ρ is, there exists always a variable t_ρ, measured by the thermal clock, with respect to which the system is in equilibrium and physics is the same as in the conventional non-relativistic statistical case!*

However, equilibrium is a state that emerges through molecular collisions; if there is no time there will be no such collisions, and no reason whatever to assume that equilibrium is the most probable state of the system. This proposal embodies a hidden assumption that time already exists.

Second, it is based on the underlying symmetries of Hamiltonian dynamics:

> *Mechanics does not single out a preferred variable, because all mechanical predictions can be obtained using the relativistic Hamiltonian H, which treats all variables on equal footing.*

However, this symmetry applies particularly to the direction of time: Hamiltonian dynamics (13.4), (13.5) has a time-reversal symmetry

$$x \to x\,, \quad t \to -t\,, \quad q \to q\,, \quad p \to -p\,, \qquad (13.43)$$

hence like every other proposal for the purely bottom-up emergence of time, it has problems determining the arrow of time. An initial reaction to any such proposal is that coarse graining from micro to macro scales convincingly results in an arrow of time, as shown beautifully by Boltzmann's H-theorem [13.51, p. 43–48], resulting from the fact that random motions in phase space takes one from less probable to more probable regions of phase space ([13.30, p. 686–696]; [13.52, p. 43–47]; [13.53, p. 9–56]). Hence one can show that entropy increases to the future; the second law of thermodynamics at the macro level emerges from the coarse grained underlying micro theory. The quantum theory version of this result is the statement that the density matrix of open system evolves in a time asymmetric manner, leading to an increase in entropy [13.54, p. 123–125].

However, this apparent appearance of an arrow of time from the underlying theory is an illusion, as the underlying theory is time symmetric, so there is no way an arrow of time can emerge by any local coarse graining procedure. Indeed the derivation of the increase of entropy in Boltzmann's H-theorem applies equally to both directions of time (swap $t \to -t$, the same derivation still holds). This is *Loschmidt's paradox* ([13.28, Fig. 7.6]; [13.30, p. 696–699]; [13.53]):

Time Symmetry of H-Theorem

Boltzmann's H-theorem predicts that entropy will increase to both the future and the past.

The same will apply to the quantum theory derivation of an increase of entropy through evolution of the density matrix ([13.54, p. 123–125], [13.52, p. 38–42], and 53–58); it cannot resolve where the arrow of time comes from, or indeed why it is the same everywhere. The latter is a key question for any local proposal for determining the arrow of time:

The Arrow of Time Locality Issue

If there is a purely local process for determining the arrow of time, why does it give the same result everywhere?

We are unaware of any contradictions with regard to the direction of the arrow of time in the universe around us, either locally (time does not run backwards anywhere on Earth) or astronomically (irreversible processes in distant galaxies seem to run in the same direction of time as here [13.55]). Some top-down coordinating mechanism is called for to guarantee that the future direction of time will be the same everywhere; that is lacking in any purely bottom-up proposal, which is by its nature based in local interactions only.

Kupervasser et al. [13.56] suggest that interaction between two subsystems with a different arrow of time will cause a decay towards a universal direction for the arrow of time. This is a very interesting claim, but has two problems: first, how can a coherent interaction take place at an interface where the direction of time is different on the two sides? It seems *a priori* that paradoxical behavior will abound, as closed causal

loops will necessarily occur there. Moreover, if this works despite these problems, then one has to show that there has been sufficient time available since the start of the universe that all domains opposite to the dominant one are coerced to join the majority; effective causal horizons [13.5] might even prevent this occurring. *Kupervasser* et al. do not put their proposal in the cosmic context, so this too is unresolved.

13.5.3 The Wheeler–de Witt Equation

Much literature on the problem of time in quantum cosmology [13.9, 57, 58] suggests that an effective time emerges from a time-independent wave function of the universe determined by the Wheeler–de Witt equation [13.59–62]. As stated by *Hartle* [13.59, 60]:

> *In quantum mechanics, any system – the universe included – is described by a wave function Ψ. There is a local dynamical law called the Schrödinger equation that governs how the wave function changes in time*
>
> $$\mathrm{i}\hbar \frac{\mathrm{d}|\Psi(t)\rangle}{\mathrm{d}t} = H|\Psi(t)\rangle \quad \textit{(dynamical law)}\,. \tag{13.44}$$
>
> *Here the operator H, the Hamiltonian, summarizes the dynamical theory .... The Schrödinger equation doesn't make any predictions itself, it requires an initial condition. This is*
>
> $$|\Psi(0)\rangle \quad \textit{(initial condition)}\,. \tag{13.45}$$
>
> *When we consider the universe as a quantum mechanical system, this initial condition is Hawking's wave function of the universe [13.61].*

This is the basis of quantum cosmology [13.63].

The problem of time then arises because in the case of general relativity, where $|\Psi\rangle$ is a function of three-geometries: $|\Psi\rangle = |\Psi(h_{ij})\rangle$, the Hamiltonian is such as to lead to the Wheeler–de Witt equation for the wave function of the universe

$$H|\Psi\rangle = 0\,. \tag{13.46}$$

So by (13.44), $|\Psi\rangle$ is time independent and the probabilities for quantum outcomes in the universe, which are expressed by the wave function $|\Psi\rangle$, are unchanging in time (this is a simplified sketch; for details, see [13.22, 23, 57]). Hence time does not pass, the universe just is. Time is an illusion [13.9]. Much literature then tries to show how an effective time can emerge from this timeless context [13.9, 57, 58]. We will make four points.

First, *Arnowitt*, *Deser*, and *Misner* write about the Hamiltonian formalism as follows [13.41]:

> *Since the relation between q_{M+1} and τ is undetermined, we are free to specify it explicitly, i. e., impose a* coordinate condition. *If, in particular, this relation is chosen to be $q_{M+1} = \tau$ (a condition which also determines N), the action (2.4) then reduces* [to] *(2.5) with the notational change $q_{M+1} \to \tau$; the nonvanishing Hamiltonian only arises as a result of this process.*

This is the choice made above (Sect. 13.4.2); the corresponding Hamiltonian will be nonzero as indicated in this quote, so (13.46) will not hold [13.47].

Second, is there an alternative proposal that does not lead to (13.46)? Yes indeed: unimodular gravity (which produces a trace free version of the Einstein field equations [13.64]) has the same effective gravitational equations as general relativity theory [13.65] but makes $H \neq 0$ and so solves the time problem of quantum cosmology. *Smolin* [13.66] states this as follows:

> *Sorkin [13.67, 68] and Unruh [13.69, 70] have pointed out that unimodular gravity has a nonvanishing Hamiltonian and hence evolves quantum states in terms of a global time given by an analogue of the Schrödinger equation.*

This removes the basis of the problem.

It has other major benefits: as emphasized by *Weinberg* [13.65] and *Smolin* [13.66], it also solves the strong cosmological constant problem; the discrepancy factor of at least 10^{70} between estimates of the vacuum energy density and the cosmologically determined value of the cosmological constant [13.20]. This resolution is a crucial need in relating quantum field theory to general relativity: it is a *sine qua non* for consistent physics, hence the following.

Reconciling General Relativity and Quantum Field Theory

Evidence from cosmology [13.20] of the small size of the cosmological constant strongly favors the trace-free version of the Einstein equations over the usual version [13.64]. (If one applies Ockham's razor (*entities must not be multiplied beyond necessity*) the proposed multiverse solution is severely disfavored in comparison with this resolution of the issue, which does not involve an infinitude of unobservable entities.)

Then we can have (13.44) without (13.46).

Third, this analysis assumes that a Hamiltonian evolution (13.44) holds for the wave function of the universe as a whole at all times. This is sometimes justified by saying that as there can be no external measuring apparatus for the universe as a whole to interact with, no measurement or wave function collapse can take place: only unitary evolution will occur. However one can advocate an emergent view of higher level laws of causation from lower level physical interactions [13.19]. On this view, the wave function of the universe $|\psi_U\rangle$ is the wave function obtained by composition of all its components: it is a sum of terms of the form

$$|\psi_U\rangle = |\psi_1\rangle \otimes |\psi_2\rangle \otimes \ldots . \otimes |\psi_N\rangle \,, \qquad (13.47)$$

where N is the number of constituents making up the universe and $|\psi_i\rangle$ the wave function for degrees of freedom of the i-th component. This will not evolve unitarily if any wave function collapse takes place for any component anywhere in the universe. However, measurements do take place and classical physics does emerge. Thus with this view, we are not entitled to assume the existence of a wave function for the entire universe that always obeys (13.46); this starting assumption is unjustified, at least in recent times.

Finally, the argument above is centrally based on the Wheeler–de Witt equation, but it is not a tested and proven part of physics.

Missing Confirmation

We have no observational or experimental evidence that the Wheeler–de Witt equation, in fact, describes the evolution of the real universe at any time.

It is an untested extrapolation of known physics, which extrapolation may or may not represent reality adequately [13.2]. Actually, one can suggest that everyday experience strongly suggests it is not true (this is hinted at in [13.22]). That is the topic of Sect. 13.6.

13.6 It's All in the Mind

Barbour [13.9] tackles the key issue of how the mind can experience the passing of time in the timeless context of the Wheeler–de Witt equation: how do we reconcile the conclusion that time is an illusion with the fact we do indeed experience the passage of time? This section reviews that suggestion, and argues that it is fatally flawed; and that this issue is a major problem for all proposals that time is an illusion.

Barbour claims [13.2, 9] that there exist records of events that our brains read sequentially, and so create a false illusion of the passage of time. Thus brain processes are responsible for illusion of change. However, *processes* are things that unfold in time – there are no processes unless time flows. One cannot perceive a flow of time unless time flows, because perception is a process that takes place in time.

The prevalent view of present-day neuroscience [13.71] is that mental states Φ are functions of brain states B, which are based in the underlying neuronal states b_i, determined by genetics, chemistry, and physics interactions in the brain, taking place in the overall physical, social, and psychological environment $\mathcal{E}$. Thus

$$\Phi = \Phi(B) = \Phi(b_i, \mathcal{E}) \,. \qquad (13.48)$$

If time does not flow in microphysics, in a given unchanging environment

$$\left\{\frac{\mathrm{d}b_i}{\mathrm{d}t} = 0, \frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} = 0\right\} \Rightarrow \mathrm{d}\Phi/\mathrm{d}t = 0 \,, \qquad (13.49)$$

mental states cannot evolve, unless they are driven in some mysterious unspecified way by changes in the environment;: but in Barbour's argument, that cannot evolve either.

However, one thing we do know is that time does flow in our experience (indeed *knowing* is a key part of that experience!). Hence the assumption that time does not flow in the underlying microphysics cannot be true; the data proves it to be wrong. If Barbour's view is correct and no physical events take place, then – as the brain is based in physics – no such record-reading processes can take place. Rather than showing time is an illusion, we suggest that the implication runs the other way.

Taking Everyday Life Seriously

Comparing the conclusion (*time is an illusion*) with evidence from mental life, by (13.49) the contradiction between them is proof that the W–deW equation (13.46) does not apply to the universe as a whole at the present time, as proposed by Barbour.

If there is a meaningful wave function of the universe (perhaps defined by (13.47)), it does not evolve

in a unitary way. The great merit of Barbour's book is that it takes the Wheeler–de Witt equation seriously and pursues the implications to their logical conclusion; the evidence from daily life then shows it to be wrong.

A further set of issues arise with regard to the perception of time. Experienced time τ_{exp} is a function of proper time τ but also of the emotional and psychological context. It also has a minimal resolution τ_{min} because of the interaction times in the brain; we cannot distinguish events at smaller times scales because our senses necessarily average over micro time scales with a window function $W(\tau_{min})$. However, all this is unrelated to the fundamental issues of the nature of time that we are dealing with here; it is to do with brain functioning. The brain is not necessarily a good clock, and it will not click over instantaneously: there will be a finite width to its time resolution. Yet the fact that it works as it does is evidence of the flow of time in physics, because the brain is based in physics. It is not just a series of correlations: it is an ordered sequence of causally related correlations that flow from each other in an ongoing process enabling mental life.

13.7 Taking Delayed Choice Quantum Effects into Account

This chapter so far has been based on a classical view of physics. The EBU proposal made so far does not take into account the delayed choice experiments of quantum theory, which suggest that one can in some circumstances *reach back into the past* to affect things there. This section briefly comments on how one can extend the EBU model to take this feature into account.

One can extend the EBU view to one that takes account of this aspect of quantum physics through proposing a *crystallizing block universe* (CBU), where *the present* is effectively the transition region in which quantum uncertainty changes to classical definiteness [13.72]. Such a crystallization, however, does not take place simultaneously, as it does in the simple classical picture. Quantum physics appears to allow some degree of influence of the present on the past, as indicated by such effects as Wheeler's delayed choice experiments [13.73, 74] and Scully's quantum eraser [13.75] (see the summaries of these effects in [13.15, 16]).

The CBU picture is an extension of the EBU where local events may lead or lag the overall flow of time, thus allowing some apparent influences from the future to the past as evidenced in those experiments. It adequately reflects such effects by distinguishing the transitional events where uncertainty changes to certainty, which may in some cases be delayed till after the apparent *present time*.

We have not here tried to relate the EBU picture to the issue of entanglement and EPR type of experiments [13.15, 16]. The way those experiments relate to simultaneity and the flow of time as proposed here is unclear; this is a topic for future research.

13.8 The Arrow of Time and Closed Time-Like Lines

Two closely related problems are the arrow of time problem and the issue of closed time-like lines. This section discusses how the EBU proposal reformulates the first issue and solves the second.

13.8.1 The Arrow of Time

With respect to the arrow of time problem [13.51, 76], if the EBU view is correct, the Wheeler–Feynman prescription for introducing the arrow of time by integration over the far future [13.77], and associated views comparing the far future with the distant past [13.28, 78, 79], are invalid approaches to solving the arrow of time problem, for it is not possible to do integrations over future time domains if they do not yet exist. Indeed, the use of half-advanced and half-retarded Feynman propagators in quantum field theory then becomes a calculational tool representing a local symmetry of the underlying physics that does not reflect the nature of emergent physical reality, in which that symmetry is broken.

The arrow of time problem in this EBU context is revisited in a companion paper [13.25]. The key point is the following.

The direction of time

The arrow of time arises fundamentally because the future does not yet exist; a global asymmetry in the physics context. The Feynman propagator can only be integrated over the past, as the future spacetime domain is yet to be determined.

One can be influenced at the present time from many causes lying in our past, as they have already taken place and their influence can be felt thereafter. One cannot be influenced by causes coming from the future, for they have not yet come into being. The history of the universe has brought the past into being, which is steadily extending to the future, and the future is just a set of unresolved potentialities at present. One cannot integrate over future events to determine their influence on the present not only because they do not yet exist, but because they are not even determined at present (Sect. 13.2).

The direction of the arrow of time is thus determined in a contingent way in the EBU context [13.25]: it is the direction of time leading from what already has come into existence (the past) to the present. Collapse of the quantum wave function is a prime candidate for a location of a physical solution to the coming-into-being problem and manifests itself as a form of time-asymmetric top-down action [13.80] from the universe as a whole to local systems [13.28]. A key further feature is that the initial state of the universe was very special with very low entropy [13.25, 53, 58], allowing complex higher entropy structures to form later on. However, that effects how the arrow of time works out, leading to the second law of thermodynamics, rather than its very existence; that is provided by the EBU context.

13.8.2 Closed Time-Like Lines: Chronology Protection

A longstanding problem for general relativity theory is that closed time-like lines can occur in exact solutions of the Einstein field equations with reasonable matter content, as shown famously in the static rotating Gödel solution [13.5]. This opens up the possibility of many paradoxes, such as killing your own grandparents before you were born and so creating causally untenable situations.

It has been hypothesized that a *chronology protection conjecture* [13.81] would prevent this happening. Various arguments have been given in its support [13.82], but this remains an *ad hoc* condition added on as an extra requirement on solutions of the field equations, which do not by themselves give the needed protection.

The EBU automatically provides such protection, because creating closed time-like lines in this context requires the undetermined part of spacetime intruding on regions that have already been fixed. But the evolving spacetime regions can never intrude into the completed past domains and so create closed time-like lines through some spacetime event P, because to do so would require the fundamental world lines to intersect each other either before reaching P or at P. Assuming plausible energy conditions, that would create a spacetime singularity [13.5], because (being time-like eigenvectors of the Ricci tensor) they are the average flow lines of matter, and in the real universe, there is always matter or radiation present: $R_{ab} \neq 0$. The extension of time cannot be continued beyond such singularities, because they are the boundary of spacetime.

Causality

The existence of closed time-like lines [13.58, p. 93–116] is prevented in an EBU, because if the fundamental world lines intersect, a spacetime singularity occurs [13.5]: the worldlines are incomplete in the future, time comes to an end there, and no *grandfather paradox* can occur.

Hence the EBU as outlined above automatically provides chronology protection.

13.9 Overall: A More Realistic View

This paper has proposed an EBU representation of spacetime which grows with time as events happen. This final section reviews how it relates to the basic features usually expected of time and to some of the surprising features of time revealed to us by relativity theory.

The EBU model recognizes that the nature of the future is completely different from the nature of the past. The past has taken place and is fixed, and so the nature of its existence is quite different than that of the indeterminate future. Uncertainty exists with respect to both the future and the past, but its nature is different in these two cases. The future is uncertain because it has not yet been determined; it does not yet exist in a physical sense (although it is constrained in key ways by the current state of things). Thus this uncertainty has an ontological character. The past, however, is fixed and unchanging, because it has already happened, and the times when it

happened cannot be revisited; but our knowledge about it is incomplete and can change with time. Thus this uncertainty is epistemological in nature.

In Newtonian theory, and in ordinary quantum theory, time is the source of:

1. Ordering of events
2. Duration measured between events
3. Simultaneity: synchronization of distant events
4. Direction of the flow of time
5. Transition: the fact that time flows
6. Continuity of the flow of time
7. The monotonic nature of that flow (it cannot reverse or close up).

However, special relativity and general relativity changed that with a surprising find [13.5]:

> ***Key discovery 1****: Simultaneity (3) is not fundamental to time: – time flows along time-like world lines, proper time along world lines is the fundamentally preferred time parameter. It is measured by the spacetime metric, which determines duration (2).*

Thus simultaneity (3) is secondary, with no direct physical consequences. What matters are interactions between distinct entities; these take place via time-like curves and null geodesics, not on space-like surfaces. This potentially puts a major barrier in the way of the EBU proposal where the flow of time is taken seriously, but this chapter has suggested that those barriers are resolved by identifying preferred time-like curves and associated space-like surfaces in realistic models of the real universe (Sect. 13.4.3).

Additionally, general relativity made a crucial difference to (7) [13.5]:

> ***Key discovery 2****: – monotonicity (7) is not necessarily true in a curved spacetime, unless something prevents it (as shown by Gödel, closed time-like lines are potentially possible even for solutions of the Einstein field equations).*

The EBU model solves this key problem (Sect. 13.8.2), which means ordering (1) is also OK in them.

Unlike the block universe models, the general relativity EBU models adequately represent (4), (5), and (6), which are the same in them as in Newtonian theory.

When quantum effects are significant, the future manifests all the signs of quantum weirdness, including duality, uncertainty, and entanglement. With the passage of time, after the time-irreversible process of state-vector reduction has taken place, the past emerges, with the previous quantum uncertainty replaced by the classical certainty of definite particle identities and states. The present time is where this transition largely takes place. However, the process does not take place uniformly or reversibly; evidence from delayed choice experiments shows that some isolated patches of quantum indeterminacy remain, and their transition from probability to certainty only takes place later. Thus, when quantum effects are significant, the EBU of classical physics cedes way to CBU[13.72]. On large enough scales that quantum effects are not significant, the two models become indistinguishable.

Interesting work to be done arising of the EBU proposal, apart from testing its basic ideas, includes

1. Determining the nature of the preferred time surfaces defined in Sect. 13.3 in inhomogeneous cosmologies (they are the same as the usual surfaces in spatially homogeneous models).
2. Extending the ADM analysis of Sect. 13.4.3 to the case where the preferred surfaces of constant time go null and then become time like.
3. Relating the geometry of time surfaces when spacetime is represented on different averaging scales; this is an aspect of the fitting and averaging problem for general relativity theory [13.83, 84].
4. Determining how the idea can sensibly relate to entanglement and EPR type experiments [13.15, 16];
5. Investigating the relation of quantum gravity theories to the EBU proposal.

We do not yet have a reliable theory of quantum gravity, but there are some proposals that do indeed see time at the quantum level as unfolding in an analogous way to the EBU (for example, spin foam models [13.85]). Our view would be that however they relate to time [13.21–23, 62, 86, 87], they must be capable of producing an EBU at the classical level or they will fail the fundamental test of relating convincingly to the physics of ordinary everyday life. This is a correspondence principle for these theories.

Conclusion. We have reviewed the many arguments against the flow of time, in particular those based in the Wheeler–de Witt equation, and have argued that they do not carry the day; the EBU is a good model of spacetime that fits well with our daily experience as well as with general relativity and quantum theory. A key issue is how the properties of time relate to the experiences we have through the operations of our mind; we have argued (Sect. 13.6) that this is crucial evidence that we must take into account.

A Key Test

The experimental evidence supporting the huge corpus of present-day neuroscience [13.71] decisively favors the EBU over the usual block universe proposal, at the classical level; therefore to be acceptable, any proposed underlying theory must pass the critical test of leading to an effective EBU at the macro level.

The physics equations we should believe are those that are compatible with this evidence; those that are not fail a basic reality test.

13.A The ADM Formalism

This Appendix further develops the relation of the proposal made here to the usual ADM formalism (Sect. 13.4.2).

Let us consider a globally hyperbolic manifold $(\mathcal{M}, g)$ having a topological structure $\Sigma \otimes \boldsymbol{R}$, where Σ_t denotes the family of space-like hypersurfaces labeled by the parameter t. On each hypersurface of constant t, we can define a purely spatial metric as $h_{ab} = g_{ab} + n_a n_b$ where n^a is the (necessarily time-like) unit normal vector of Σ_t with $n^a n_a = -1$. Hence, given the foliation on $(\mathcal{M}, g)$, the spatial metric h_{ab} on the space-like hypersurface Σ_t is uniquely defined.

Let us also assume that the Ricci tensor R_{ab} on this manifold has one time-like and three space-like eigenvectors. These eigenvectors are unique for any physically realistic (*Type I*) nonzero matter field. Let the timeline for a given observer, be the integral curve of the time-like eigenvector t^a of R_{ab}. This then uniquely defines the *shift*-vector with respect to a given foliation of a family of space-like hypersurfaces Σ_t as

$$N^a = h^a_b t^b \, . \tag{13.50}$$

Furthermore, if we specify the relation between this coordinate time t and proper time τ as $\mathrm{d}\tau = N(t, x^i)\,\mathrm{d}t$ (where x^i are the coordinates on the three-surface Σ_t), then by definition this gives the lapse function

$$N(t, x^i) = -t^a n_a \, . \tag{13.51}$$

Specifically if the coordinate time is equal to the proper time then we must have $t^a n_a = -1$.

References

13.1 P.C.W. Davies: That mysterious flow, Sci. Am. **21**, 8–13 (2012), special edition

13.2 J.N. Butterfield: The End of Time?, arXiv:gr-qc/0103055 (2010)

13.3 FQXI essay competition: http://fqxi.org/community/forum/category/10 (2010)

13.4 FQXI meeting on time: http://fqxi.org/conference/2011 (2011)

13.5 S.W. Hawking, G.F.R. Ellis: *The Large Scale Structure of Space-Time* (Cambridge Univ. Press, Cambridge 1973)

13.6 D.H. Mellor: *Real Time II* (Routledge, London 1998)

13.7 S. Savitt: Being and becoming in modern physics. In: The Stanford Encyclopedia of Philosophy (Spring 2002 Edition), ed. by E.N. Zalta, available online at http://plato.stanford.edu/archives/spr2002/entries/spacetime-bebecome/

13.8 H. Price: *Time's Arrow and Archimedes' Point* (Oxford Univ. Press, New York 1996)

13.9 J.B. Barbour: *The End of Time: The Next Revolution in Physics* (Oxford Univ. Press, Oxford 1999)

13.10 C. Rovelli: Forget time, FQXI essay (2008), available online at http://fqxi.org/community/forum/topic/237

13.11 C.D. Broad: *Scientific Thought* (Harcourt Brace, New York 1923), for table of contents and some chapters see http://www.ditext.com/broad/st/st-con.html

13.12 G.F.R. Ellis: Physics in the real universe: Time and spacetime, GRG **38**, 1797–1824 (2006), arXiv:gr-qc/0605049

13.13 I. Kanter, Y. Aviad, I. Reidler, E. Cohen, M. Rosenbluh: An optical ultrafast random bit generator, Nat. Photonics **4**, 58–61 (2010)

13.14 R.P. Feynman, R.B. Leighton, M. Sands: *The Feynman Lectures on Physics: Quantum Mechanics* (Addison-Wesley, Reading 1965)

13.15 Y. Aharanov, D. Rohrlich: *Quantum Paradoxes. Quantum Theory for the Perplexed* (Wiley-VCH, Weinheim 2005)

13.16 G. Greenstein, A.G. Zajonc: *The Quantum Challenge: Modern Research on the Foundations of*

Quantum Mechanics (Jones and Bartlett, Sudbury 2006)
13.17 S. Caroll: Ten things everyone should know about time, Discover Magazine, Kalmbach Publishing Co. (2011), available online at http://blogs.discovermagazine.com/cosmicvariance/2011/09/01/ten-things-everyone-should-know-about-time/
13.18 A.S. dington (Ed.): *The Nature of the Physical World* (MacMillan, London 1928)
13.19 G.F.R. Ellis: On the limits of quantum theory: contextuality and the quantum-classical cut, Ann. Phys. **327**, 1890–1932 (2012), arXiv:1108.5261
13.20 S. Dodelson: *Modern Cosmology* (Academic, New York 2003)
13.21 C.J. Isham: Canonical quantum gravity and the problem of time, Lectures at the NATO Summer School held in Salamanca (1992), gr-qc/9210011
13.22 N. Huggett, T. Vistarini, C. Wuthrich: Time in quantum gravity. In: *The Blackwell Companion to the Philosophy of Time*, ed. by A. Bardon, H. Dyke (Wiley-Blackwell, Chichester 2012), arXiv:1207.1635
13.23 E. Anderson: Problem of time in quantum gravity, arXiv:1206.2403 (2012)
13.24 R.P. Feynman, R.B. Leighton, M. Sands: *The Feynman Lectures on Physics: Mainly Mechanics, Radiation, and Heat* (Addison-Wesley, Reading 1963)
13.25 G.F.R. Ellis: The arrow of time, the nature of spacetime, and quantum measurement (2011), available online at http://www.mth.uct.ac.za/~ellis/Quantum_arrowoftime_gfre.pdf
13.26 M.R. Spiegel: *Theory and Problems of Theoretical Mechanics* (Schaum/McGraw-Hill, New York 1967)
13.27 R. Feynman: *QED: The Strange Theory of Light and Matter* (Princeton Univ. Press, Princeton 1985)
13.28 R. Penrose: *The Emperor's New Mind* (Oxford Univ. Press, Oxford 1989)
13.29 C.J. Isham: *Lectures on Quantum Theory: Mathematical and Structural Foundations* (Imperial College Press, London 1997)
13.30 R. Penrose: *The Road to Reality: A complete guide to the Laws of the Universe* (Jonathan Cape, London 2004)
13.31 I. Percival: Schrödinger's quantum cat, Nature **351**, 357 (1991)
13.32 J. Scalo, J.C. Wheeler, P. Williams: Intermittent jolts of galactic UV radiation: Mutagenetic effects, Frontiers of Life. 12th Rencontres de Blois, ed. by L.M. Celnikier (2001), astro-ph/0104209
13.33 G.F.R. Ellis, R.M. Williams: *Flat and Curved Space Times*, 2nd edn. (Oxford Univ. Press, Oxford 2000)
13.34 G.F.R. Ellis: Relativistic cosmology, General Relativity and Cosmology, Proc. Int. School Phys. "Enrico Fermi" (Varenna), Course XLVII, ed. by R.K. Sachs (Academic, Elsevier 1971) pp. 104–179, Reprinted as Golden Oldie: Gen. Relativ. Gravit. **41**, 581 (2009)
13.35 P.W. Anderson: More is different, Science **177**, 393–396 (1972)
13.36 G.F.R. Ellis, D.R. Matravers: Spatial Homogeneity and the size of the universe. In: *A Random Walk in Relativity and Cosmology (Raychaudhuri Festschrift)*, ed. by N. Dadhich, J.K. Rao, J.V. Narlikar, C.V. Vishveshswara (Wiley Eastern, Delhi 1985) pp. 92–108
13.37 R.P. Feynman, R.B. Leighton, M. Sands: *The Feynman Lectures on Physics: The Electromagnetic Field* (Addison-Wesley, Reading 1964)
13.38 D. Sobel: *Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time* (Walker and Company, New York 1995)
13.39 K.S. Thorne: *Black Holes and Time Warps: Einstein's Outrageous Legacy* (W. W. Norton, New York 1995)
13.40 G.F.R. Ellis: The dynamics of pressure-free matter in general relativity, J. Math. Phys. **8**, 1171–1194 (1967)
13.41 R. Arnowitt, S. Deser, C.W. Misner: The dynamics of general relativity. In: *Gravitation: An Introduction to Current Research*, ed. by L. Witten (Wiley, New York 1962) pp. 227–265, Reprinted in Gen. Relativ. Gravit. **40**, 1997 (2008)
13.42 A.R. King, G.F.R. Ellis: Tilted homogeneous cosmologies. Comm.un Math, Phys. **31**, 209–242 (1973)
13.43 P. Anninos: Computational cosmology: From the early universe to the large scale structure, Living Rev, Relativ. **4**, 2 (2001)
13.44 J.V. Narlikar: *The Lighter Side of Gravity* (Cambridge Univ. Press, Cambridge 1996)
13.45 M. Haahr: Introduction to Randomness and Random Numbers, (1999), available online athttp://www.random.org/randomness/
13.46 R. Gambini, J. Pullin: *A First Course in Loop Quantum Gravity* (Oxford Univ. Press, Oxford 2012)
13.47 A. Peres: Critique of the Wheeler-De Witt equation. In: *On Einstein's Path*, ed. by A. Harvey (Springer, New York 1998) pp. 367–379, arXiv:gr-qc/9704061v2 (1997)
13.48 N.F. Mott: Time dependence in quantum mechanics, Proc. Camb. Phil. Soc. **27**, 553 (1931)
13.49 J.S. Briggs, J.M. Rost: Time dependence in quantum mechanics, Eur. Phys. J. D **10**, 311–318 (2000)
13.50 J.S. Briggs, J.M. Rost: On the derivation of the time-dependent equation of Schrödinger, Found. Phys. **31**, 693–712 (2001)
13.51 H.-D. Zeh: *The Physical Basis of the Direction of Time* (Springer, Berlin, Heidelberg 2007)
13.52 J. Gemmer, M. Michel, G. Mahler: *Quantum Thermodynamics: Emergence of Thermodynamic Behaviour Within Composite Quantum Systems* (Springer, Heidelberg 2004)
13.53 R. Penrose: *Cycles of Time: An Extraordinary New View of the Universe* (Knopf, New York 2011)
13.54 H.-P. Breuer, F. Petruccione: *The Theory of Open Quantum Systems* (Clarendon Press, Oxford 2006)
13.55 M.J. Rees: *Perspectives in Astrophysical Cosmology* (Cambridge Univ. Press, Cambridge 1995)

13.56 O. Kupervasser, H. Nikoli, V. Zlati: The universal arrow of time (2012) arXiv:1011.4173
13.57 J. Halliwell: The interpretation of quantum cosmology and the problem of time. In: *The Future of Theoretical Phyics and Cosmology: Celebrating Stephen Hawking's 60th Birthday*, ed. by G.W. Gibbons, E.P.S. Shellard, S.J. Rankin (Cambridge Univ. Press, Cambridge 2003) pp. 675–690
13.58 S. Carroll: *From Eternity to Here: The Quest for the Ultimate Arrow of Time* (Dutton, New York 2010)
13.59 J. Hartle: Theories of everything and Hawking's wave function. In: *The Future of Theoretical Physics and Cosmology: Celebrating Stephen Hawking's 60th Birthday*, ed. by G.W. Gibbons, E.P.S. Shellard, S.J. Rankin (Cambridge Univ. Press, Cambridge 2003) pp. 38–49
13.60 J. Hartle: The state of the universe. In: *The Future of Theoretical Phyics and Cosmology: Celebrating Stephen Hawking's 60th Birthday*, ed. by G.W. Gibbons, E.P.S. Shellard, S.J. Rankin (Cambridge Univ. Press, Cambridge 2003) pp. 615–620
13.61 S.W. Hawking: The quantum state of the universe, Nucl. Phys. **B239**, 2447 (1984)
13.62 J. Butterfield, C.J. Isham: On the emergence of time in quantum gravity. In: *The Arguments of Time*, ed. by J. Butterfield (Oxford, Oxford Univ. Press 1999), arXiv:gr-qc/9901024v1
13.63 G.W. Gibbons, E.P.S. Shellard, S.J. Rankin (Eds.): *The Future of Theoretical Physics and Cosmology: Celebrating Stephen Hawking's 60th Birthday* (Cambridge Univ. Press, Cambridge 2003)
13.64 G.F.R. Ellis, H. van Elst, J. Murugan, J.-P. Uzan: On the trace-free Einstein equations as a viable alternative to general relativity, Class. Quantum Gravity **28**, 225007 (2011), arXiv:1008.1196
13.65 S. Weinberg: The cosmological constant problem, Rev. Mod. Phys. **61**, 1–23 (1989)
13.66 L. Smolin: Quantization of unimodular gravity and the cosmological constant problems, Phys. Rev. D **80**, 084003 (2009), arXiv:0904.4841v1 [hep-th
13.67 R.D. Sorkin: On the role of time in the sum-over-histories framework for gravity int, J. Theor. Phys. **33**, 523–534 (1994)
13.68 R.D. Sorkin: Spacetime and causal sets (1991), available online at http://www.cdms.syr.edu/~sorkin/some.papers/66.cocoyoc.pdf
13.69 W.G. Unruh: A unimodular theory of canonical quantum gravity, Phys. Rev. D **40**, 1048 (1989)
13.70 W.G. Unruh, R.M. Wald: Time and the interpretation of canonical quantum gravity, Phys. Rev. D **40**, 2598 (1989)
13.71 E.R. Kandel, J.H. Schwartz, T.M. Jessell: *Principles of Neuroscience* (McGraw Hill, New York 2000)
13.72 G.F.R. Ellis, T. Rothman: Crystallizing block universes, Int. J. Theor. Phys. **49**, 988 (2010), arXiv:0912.0808
13.73 J.A. Wheeler: The "past" and the "delayed-choice double-slit experiment". In: *Mathematical Foundations of Quantum Theory*, ed. by A.R. Marlow (Academic, PLARV 1978) pp. 9–48
13.74 V. Jacques, E. Wu, F. Grosshans, F. Treussart, P. Grangier, A. Aspect, J.-F. Roch: Experimental realization of Wheeler's delayed-choice Gedanken-Experiment, Science **315**, 5814 (2007), arXiv:quant-ph/0610241v1
13.75 Y.-H. Kim, R. Yu, S.P. Kulik, Y.H. Shih, M.O. Scully: A Delayed Choice Quantum Eraser, Phys. Rev. Lett. **84**, 1–5 (2000), arXiv:quant-ph/9903047v1
13.76 P.C.W. Davies: *The Physics of Time Asymmetry* (Surrey Univ. Press, London 1974)
13.77 J.A. Wheeler, R.P. Feynman: Interaction with the absorber as the mechanism of radiation, Rev. Mod. Phys. **17**, 157–181 (1945)
13.78 G.F.R. Ellis, D.W. Sciama: Global and non-global problems in cosmology. In: *General Relativity (A. Synge Festschrift)*, ed. by L. O'Raifeartaigh (Oxford Univ. Press, Oxford 1972) pp. 35–59
13.79 G.F.R. Ellis: Cosmology and local physics, New Astron. Rev. **46**, 645–658 (2002), gr-qc/0102017)
13.80 G.F.R. Ellis: On the nature of causation in complex systems, Trans. R. Soc. South Africa **63**, 69–84 (2008)
13.81 S.W. Hawking: The chronology protection conjecture, Phys. Rev. D **46**, 603–611 (1992)
13.82 M. Visser: The quantum physics of chronology protection. In: *The Future of Theoretical Physics and Cosmology: Celebrating Stephen Hawking's 60th Birthday*, ed. by G.W. Gibbons, E.P.S. Shellard, S.J. Rankin (Cambridge Univ. Press, Cambridge 2002) pp. 161–173, arXiv:gr-qc/0204022v2
13.83 G.F.R. Ellis: Relativistic cosmology: Its nature, aims and problems. In: *General Relativity and Gravitation*, ed. by B. Bertotti, F. de Felice, A. Pascolini (Reidel, Netherlands 1984) pp. 215–288
13.84 G.F.R. Ellis, W.R. Stoeger: The fitting problem in cosmology, Class. Quantum Gravity **4**, 1679–1690 (1987)
13.85 J.C. Baez: Spin foam models, Class. Quantum Gravity **15**, 1827–1858 (1998), arXiv:gr-qc/9709052v3
13.86 C.J. Isham: Prima facie questions in quantum gravity (1993) arXiv:gr-qc/9310031v1
13.87 J. Butterfield, C.J. Isham: Spacetime and the philosophical challenge of quantum gravity. In: *Physics meets Philosophy at the Planck Scale*, ed. by C. Callender, N. Huggett (Cambridge Univ. Press, Cambridge 2000), arXiv:gr-qc/9903072v1

14. Unitary Representations of the Inhomogeneous Lorentz Group and Their Significance in Quantum Physics

Norbert Straumann

Minkowski's great discovery of the spacetime structure behind Einstein's special theory of relativity (SR) had an enormous impact on much of twentieth-century physics. (For a historical account of Minkowski's *Raum und Zeit* lecture and Poincaré's pioneering contribution, we refer to [14.1] and Chap. 2.) The symmetry requirement of physical theories with respect to the automorphism group of Minkowski spacetime – the inhomogeneous Lorentz or Poincaré group – is particularly constraining in the domain of relativistic quantum theory and led to profound insights. Among the most outstanding early contributions are Wigner's great papers on relativistic invariance [14.2]. His description of the (projective) irreducible representations of the inhomogeneous Lorentz group, that classified single particle states in terms of mass and spin, has later been taken up on the mathematical side by *George Mackey*, who developed Wigner's ideas into a powerful theory with a variety of important applications [14.3–5]. Mackey's theory of induced representations has become an important part of representation theory for locally compact groups. For certain classes it provides a full description of all irreducible unitary representations.

We find it rather astonishing that this important classical subject is not treated anymore in most modern textbooks on quantum field theory.

I shall begin with general remarks on symmetries in quantum theory, and then repeat Wigner's heuristic analysis of the unitary representations of the homogeneous Lorentz group (more precisely, of the universal covering group of the one-component of that group). This will lead us to those parts of Mackey's theory of induced representations which are particularly useful for physicists. In the final section, we shall describe free classical and quantum fields for arbitrary spin, and show that locality implies the normal spin-statistics connection. We shall see that the theory of free fields is a straightforward application of Wigner's representations of the inhomogeneous Lorentz group. (Since the quantum theory for massless fields poses delicate problems – as is well known for spin 1 – we treat only the massive case.)

14.1 Lorentz Invariance in Quantum Theory

In this section, we recall why the requirement of the restricted Lorentz invariance in quantum theory can be described in terms of unitary representations of the universal covering group of the one-component of the Poincaré group $\mathcal{P}_+^\uparrow$.

14.1.1 Symmetry Operations in Quantum Theory

In quantum theory, a symmetry operation is realized by a *Wigner automorphism*, that is by a bijection α of the set of unit rays of the underlying Hilbert space $\mathcal{H}$ (the projective space $\mathcal{P}(\mathcal{H})$ of $\mathcal{H}$), which satisfies the invariance property

$$\langle \alpha([\phi]), \alpha([\psi]) \rangle = \langle [\phi], [\psi] \rangle \,, \tag{14.1}$$

where the scalar product of two unit rays $[\phi], [\psi]$ is defined by $\langle [\phi], [\psi] \rangle = |\langle \phi, \psi \rangle|$, with $\phi \in [\phi]$, $\psi \in [\psi]$. A well-known theorem of Wigner states that every Wigner automorphism is induced by a unitary or antiunitary transformation, i. e., α is of the form

$$\alpha([\psi]) = [U\psi] \,, \quad \psi \in [\psi] \,, \tag{14.2}$$

where U is either unitary or antiunitary, and is uniquely determined up to an overall phase. (In this section, we quote various profound facts. For references to proofs, see e.g. [14.6].)

14.1.2 Projective and Unitary Representations

A symmetry group G is represented by Wigner automorphisms α_g, $g \in G$, satisfying

$$\alpha_{g_1} \circ \alpha_{g_2} = \alpha_{g_1 g_2} \,. \tag{14.3}$$

We say that $g \mapsto \alpha_g$ is a *projective representation* of G. By Wigner's theorem each α_g is induced by a unitary or antiunitary transformation U_g, which is unique up to a phase factor. For any choice we obtain from (14.3)

$$U_{g_1} U_{g_2} = \omega(g_1, g_2) U_{g_1 g_2} \,, \quad |\omega(g_1, g_2)| = 1 \,. \tag{14.4}$$

Let us now consider topological groups, especially Lie groups, and require that $g \mapsto \alpha_g$ is weakly continuous. This means that $g \mapsto \langle [\chi], \alpha_g([\phi]) \rangle$ is a continuous function for all $[\chi], [\phi] \in \mathcal{P}(\mathcal{H})$. Each U_g for g in the one-component G^0 of G is then unitary if G^0 is a Lie group. First of all, each element in a sufficiently small neighborhood $\mathcal{N}(e)$ of the unit element e can be represented as a square: for $a = \exp(X) \in \mathcal{N}(e)$ we have $a = b^2$, $b = \exp(X/2) \in \mathcal{N}(e)$, hence U_a is unitary. Now, each $g \in G^0$ can be represented as a finite product $g = a_1 \ldots a_n$, with $a_k \in \mathcal{N}(e)$. This proves the claim.

The following theorem of Bargmann is central.

Theorem 14.1 Bargmann
The phase freedom can be used such that in a some neighborhood $\mathcal{N}(e)$ the map $g \mapsto U_g$ is strongly continuous.

Can one use the remaining phase freedom such that the multipliers $\omega(g_1, g_2)$ are at least locally equal to 1? The following is true:

Theorem 14.2 Bargmann
In a sufficiently small neighborhood of e, the choice $\omega(g_1, g_2) \equiv 1$ is possible for semisimple Lie groups (such as $SO(n), L_+^\uparrow$) and affine linear groups, in particular $\mathcal{P}_+^\uparrow$. More precisely, this is exactly the case when the second cohomology group $H^2(\mathcal{G}, \mathbb{R})$ of the Lie algebra $\mathcal{G}$ of G is trivial.

Remark 14.1
It is physically significant that this is not possible for the Galilei group.

In this situation, we have a *local* strongly continuous unitary representation of G^0: $U_{g_1} U_{g_2} = U_{g_1 g_2}$. If G^0 is not simply connected, there is no reason that the multipliers $\omega(g_1, g_2)$ can be transformed away glob-

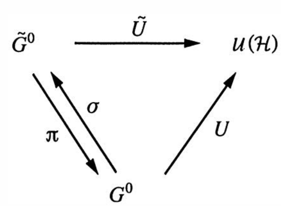


Fig. 14.1 Diagram for the lifted representation

ally. This becomes, however, possible if we pass to the universal covering group $\tilde{G}^0$ of G^0. These groups differ globally as follows: if $\pi : \tilde{G}^0 \to G^0$ is the covering map, the kernel N of π is a discrete central normal subgroup.

Now, the local representation of G^0 induces via the local isomorphism with $\tilde{G}^0$ a local representation of the universal covering group $\tilde{G}^0$. Since this group is simply connected, there is a unique extension to a strongly continuous unitary representation $\tilde{U}$ of $\tilde{G}^0$. This is indicated in Fig. 14.1, in which $\mathcal{U}(\mathcal{H})$ denotes the set of unitary operators of the Hilbert space $\mathcal{H}$.

The liftet representation $\tilde{U}_{\tilde{g}}$ of $\tilde{G}^0$ has the property

$$\tilde{U}_{\tilde{g}} = \lambda \mathbf{1}\,, \quad |\lambda| = 1 \quad \text{for } \tilde{g} \in N = \ker(\pi)\,. \tag{14.5}$$

Conversely, a representation $\tilde{U} : \tilde{G}^0 \to \mathcal{U}(\mathcal{H})$, satisfying the property (14.5), induces a projective representation of G^0. For this, choose a section $\sigma : G^0 \to \tilde{G}^0$ with $\pi \circ \sigma = id_{G^0}$ and set $U_g := \tilde{U}_{\sigma(g)}$. Since $\sigma(g_1)\sigma(g_2)$ and $\sigma(g_1 g_2)$ are in the same coset of $\tilde{G}^0/N$, the map $g \mapsto U_g$ is indeed a projective representation.

In particular, projective representations U of $\mathcal{P}_+^\uparrow$ are in one-to-one correspondence with unitary representations $\tilde{U}$ of its universal covering group $\tilde{\mathcal{P}}_+^\uparrow$ that satisfy the condition $\tilde{U}_{-e} = \pm\mathbf{1}$.

At this point, we recall the concrete form of $\tilde{\mathcal{P}}_+^\uparrow$. The universal covering group of $L_+^\uparrow$ is $SL(2,\mathbb{C})$. The twofold covering homomorphism $\lambda : SL(2,\mathbb{C}) \longrightarrow L_+^\uparrow$ is determined as follows

$$\underline{\lambda(A)x} = A\underline{x}A^\dagger\,, \tag{14.6}$$

where $\underline{x}$ denotes for each $x \in \mathbb{R}^4$ the Hermitian 2×2 matrix

$$\underline{x} = x^\mu \sigma_\mu\,, \quad \sigma_\mu = (\mathbf{1}, \sigma_k)\,. \tag{14.7}$$

(Here σ_k are the Pauli matrices, and $A^\dagger$ denotes the Hermitian conjugate of A.) From

$$\underline{x} = \begin{pmatrix} x^0 + x^3 & x^1 - \mathrm{i}x^2 \\ x^1 + \mathrm{i}x^2 & x^0 - x^3 \end{pmatrix}, \tag{14.8}$$

it follows that

$$\det \underline{x} = xx\,, \quad xy = \eta_{\mu\nu} x^\mu y^\nu = x^T \eta y\,,$$
$$\eta = (\eta_{\mu\nu}) = \mathrm{diag}(1,-1,-1,-1)\,. \tag{14.9}$$

Using this it is easy to see that the assignment $A \longmapsto \lambda(A)$ is a homomorphism from $SL(2,\mathbb{C})$ into $L_+^\uparrow$. One can show that the image is all of $L_+^\uparrow$ [14.7, 8].

The universal covering group of $\mathcal{P}_+^\uparrow$ is the semidirect product $\mathbb{R}^4 \rtimes SL(2,\mathbb{C})$, where the action of $SL(2,\mathbb{C})$ is given by $a \in \mathbb{R}^4 \mapsto \lambda(A)a$. The covering homomorphism is $(a,A) \mapsto (a, \lambda(A))$.

We assume that the reader is familiar with the spinor calculus and the finite-dimensional representations of $SL(2,\mathbb{C})$ (see the cited references).

14.2 Wigner's Heuristic Derivation of the Projective Representations of the Inhomogeneous Lorentz Group

In this section we give, following Wigner, a physicist way of arriving at the unitary irreducible representations of $\tilde{\mathcal{P}}^0 \equiv \mathbb{R}^4 \rtimes SL(2,\mathbb{C})$. A rigorous treatment has been given by G. Mackey (Sect. 14.3).

Let $(a,A) \mapsto U(a,A)$ be a unitary representation of $\tilde{\mathcal{P}}^0$ in a Hilbert space $\mathcal{H}$. If we restrict this representation to the subgroup of translations $(a,\mathbf{1})$, we get a unitary representation $U(a)$ of the translation group. According to a generalization of Stone's theorem (SNAG (Stone–Naimark–Ambrose–Godement) theorem), $U(a)$ has the representation

$$U(a) = \mathrm{e}^{\mathrm{i}Pa}\,, \tag{14.10}$$

where P^μ are commuting self-adjoint operators, interpreted as energy–momentum operators. The support of their spectral measure is Lorentz invariant. Since they commute we can choose an improper basis of eigenstates of P_μ

$$P_\mu |p,\lambda\rangle = p_\mu |p,\lambda\rangle\,, \tag{14.11}$$

where λ is a degeneracy parameter, to be determined later. (Working with improper states is, of course, formal.) We choose the covariant normalization

$$\langle p', \lambda' | p, \lambda\rangle = \delta_{\lambda'\lambda} 2p^0 \delta^{(3)}(\mathbf{p}' - \mathbf{p})\,.$$

Note that

$$U(a)|p,\lambda\rangle = \mathrm{e}^{\mathrm{i}pa}|p,\lambda\rangle\,. \tag{14.12}$$

14.2.1 Positive Mass Representations

Let us first consider the case when the momenta are on a positive mass hyperboloid $H_m^+ = \{p|p^2 = m^2, p^0 > 0\}$. Consider the standard momentum $\pi = (m, \mathbf{0})$ on this $SL(2,\mathbb{C})$ invariant orbit in momentum space, and introduce for each $p \in H_m^+$ an $SL(2,\mathbb{C})$ transformation $L(p)$ with the property $L(p)\pi = p$ ($L(p)q$ is an abbreviation for $\lambda(L(p))q$). So, $L(p)\underline{\pi}L^\dagger(p) = \underline{p}$, thus, since $\underline{\pi} = m\mathbf{1}$,

$$L(p)L^\dagger(p) = \frac{\underline{p}}{m}\,. \tag{14.13}$$

Various convenient choices of the map $p \mapsto L(p)$ will be introduced later.

Now we consider the state $U(L(p))|\pi,\lambda\rangle$. This has momentum p because

$$\begin{aligned}U(a)U(L(p))|\pi,\lambda\rangle &= U(L(p))U(L(p)^{-1}a)|\pi,\lambda\rangle\\ &= \exp(iL(p)^{-1}a\pi)U(L(p))|\pi,\lambda\rangle\\ &= \mathrm{e}^{ipa}U(L(p))|\pi,\lambda\rangle\,.\end{aligned}$$

We choose the degeneracy parameter λ for an arbitrary p such that

$$|p,\lambda\rangle = U(L(p))|\pi,\lambda\rangle\,. \tag{14.14}$$

The vectors $|\pi,\lambda\rangle$ are transformed under $SU(2)$ among themselves, because for $R \in SU(2)$

$$U(a)U(R)|\pi,\lambda\rangle = \mathrm{e}^{\mathrm{i}\pi a}U(R)|\pi,\lambda\rangle\,.$$

$SU(2)$ is the little (stability) group of π. Hence, the subspace spanned by $|\pi,\lambda\rangle$ carries a representation D of $SU(2)$

$$U(R)|\pi,\lambda\rangle = \sum_{\lambda'}|\pi,\lambda'\rangle D_{\lambda'\lambda}(R)\,. \tag{14.15}$$

For an arbitrary $A \in SL(2,C)$ we can write

$$A = L(\Lambda_A p)W(p,A)L(p)^{-1}\,, \tag{14.16}$$

where $\Lambda_A \equiv \lambda(A)$ and

$$W(p,A) := L(\Lambda_A p)^{-1}AL(p)\,. \tag{14.17}$$

One can easily see that $W(p,A)$ is an element of the little group of π. This is a so-called *Wigner rotation*. Using this decomposition, we obtain

$$\begin{aligned}U(A)|p,\lambda\rangle &= U(L(\Lambda_A p))U(W(p,A))|\pi,\lambda\rangle\\ &= \sum_{\lambda'}|\Lambda_A p,\lambda'\rangle D_{\lambda'\lambda}(W(p,A))\,.\end{aligned}$$

This explicitly shows that for an irreducible representation of $\tilde{\mathcal{P}}^0$, the representation $R \mapsto D(R)$, $R \in SU(2)$ of the little group $SU(2)$ has to be irreducible. Furthermore, only states with momenta in the orbit H_m^+ are transformed among themselves. If we choose for the irreducible representations $D^{(s)}$, $s = 0, 1/2, 1, \ldots$, the usual canonical basis, we find the following result

$$\begin{aligned}U(A)|p,\lambda\rangle &= \sum_{\lambda'}|\Lambda_A p,\lambda'\rangle D^{(s)}_{\lambda'\lambda}(W(p,A))\,,\\ W(p,A) &= L(\Lambda_A p)^{-1}AL(p)\,,\\ U(a)|p,\lambda\rangle &= \mathrm{e}^{ipa}|p,\lambda\rangle\,.\end{aligned} \tag{14.18}$$

Reformulation

Up to now we have worked with improper states $|p,\lambda\rangle$. We now translate our result to a mathematically proper formulation.

Consider superpositions

$$|\psi\rangle = \sum_\lambda \int_{H_m^+} \mathrm{d}\Omega_m(p)f_\lambda(p)|p,\lambda\rangle\,,$$

where $\mathrm{d}\Omega_m$ is the Lorentz invariant measure

$$\mathrm{d}\Omega_m(p) = \frac{\mathrm{d}^3 p}{2p^0}\,,\quad p^0 = \sqrt{\mathbf{p}^2 + m^2}\,.$$

On this we apply $U(a,A) = U(a)U(A)$ and proceed formally

$$\begin{aligned}U(a,A)|\psi\rangle &= \sum_{\lambda',\lambda}\int \mathrm{d}\Omega_m(p)f_\lambda(p)\mathrm{e}^{\mathrm{i}\Lambda_A pa}\\ &\quad\times D^{(s)}_{\lambda'\lambda}(W(p,A))\,|\Lambda_A p,\lambda'\rangle\\ &= \sum_{\lambda',\lambda}\int \mathrm{d}\Omega_m(p)f_{\lambda'}\left(\Lambda_A^{-1}p\right)\mathrm{e}^{ipa}\\ &\quad\times D^{(s)}_{\lambda\lambda'}\left(W\left(\Lambda_A^{-1}p,A\right)\right)|p,\lambda\rangle.\end{aligned}$$

Hence, the transformation of the functions $f_\lambda(p)$ is given by

$$\begin{aligned}(U^{(m,s)}(a,A)f)_\lambda(p) = \mathrm{e}^{ipa}\sum_{\lambda'}D^{(s)}_{\lambda\lambda'}(R(p,A))\\ \times f_{\lambda'}\left(\Lambda_A^{-1}p\right)\,,\end{aligned} \tag{14.19}$$

where

$$\begin{aligned}R(p,A) &= W\left(\Lambda_A^{-1}p,A\right)\\ &= L(p)^{-1}AL\left(\Lambda_A^{-1}p\right) \in SU(2)\,.\end{aligned} \tag{14.20}$$

This is a unitary representation in the Hilbert space $\mathcal{H}(m,s) = L^2(H_m^+, \mathrm{d}\Omega_m; \mathbb{C}^{2s+1})$, with the scalar product

$$\langle f, g\rangle = \sum_\lambda \int_{H_m^+} \mathrm{d}\Omega_m(p)\bar{f}_\lambda(p) g_\lambda(p)\,.$$

One can show that this representation, which is now mathematically well defined, is irreducible. It describes, in the terminology of Wigner, elementary systems with mass m and spin s.

Two Choices for the Boosts $L(p)$. As a first possibility we choose the positive Hermitian solution of (14.13), corresponding to a special Lorentz transformation in the **p**-direction. This $L(p)$ is given by

$$L(p) = \frac{1}{m^{1/2}}(\underline{p})^{1/2} = \frac{m+\underline{p}}{\sqrt{2m(m+p^0)}}\,. \tag{14.21}$$

A second choice, which leads to helicity states, uses the polar decomposition

$$L(p) = R(p)H(p)\,, \quad R(p) \in SU(2)\,,$$
$$H(p) \text{ positive Hermitian}\,.$$

$H(p)$ leads to a special Lorentz transformation in the z-direction that carries π into $(p^0, 0, 0, |\mathbf{p}|)$, and $R(p)$ rotates the z-direction into the **p**-direction. Explicitly,

$$H(p) = \begin{pmatrix} \sqrt{\frac{p^0+|\mathbf{p}|}{m}} & 0 \\ 0 & \sqrt{\frac{p^0-|\mathbf{p}|}{m}}\,, \end{pmatrix} \tag{14.22}$$

and $R(p) = \mathrm{e}^{-\mathrm{i}(\varphi/2)\sigma_3}\mathrm{e}^{-\mathrm{i}(\vartheta/2)\sigma_2}$, where ϑ, φ are the polar angles of the 3-momentum.Thus,

$$R(p) = \begin{pmatrix} \mathrm{e}^{-\mathrm{i}\varphi/2}\cos\frac{\vartheta}{2} & -\mathrm{e}^{-\mathrm{i}\varphi/2}\sin\frac{\vartheta}{2} \\ \mathrm{e}^{\mathrm{i}\varphi/2}\sin\frac{\vartheta}{2} & \mathrm{e}^{\mathrm{i}\varphi/2}\cos\frac{\vartheta}{2} \end{pmatrix}. \tag{14.23}$$

For the physical meaning of the degeneracy parameter λ, let J_k, $k = 1, 2, 3$ be the infinitesimal generators of the rotations about the x_k-axis. We interpret these as (total) angular momentum operators. Now,

$$U(L(p))J_3U^{-1}(L(p)) = U(R(p))J_3U^{-1}(R(p))$$
$$= \mathbf{J}\hat{\mathbf{p}}\,,$$

where $\hat{\mathbf{p}} = \mathbf{p}/|\mathbf{p}|$. The first equation holds because the special Lorentz transformation in the z-direction commutes with the rotations about the z-axis. From this we conclude

$$\mathbf{J}\hat{\mathbf{p}}|p,\lambda\rangle = \mathbf{J}\hat{\mathbf{p}}U(L(p))|\pi,\lambda\rangle$$
$$= U(L(p))J_3|\pi,\lambda\rangle = \lambda|p,\lambda\rangle\,.$$

Hence the parameter λ is the helicity and $|p,\lambda\rangle$ are the *helicity eigenstates.*

14.2.2 Massless Representations

Among the additional orbits we consider only the forward light cone $V^\uparrow = \{p|p^2 = 0, p^0 > 0\}$ (without the origin). The method is the same as for $m > 0$. As standard vector of the orbit we take $\pi = (1/2, 0, 0, 1/2)$. The boosts $L(p)$ still satisfy (14.13), and the degeneracy parameters λ are again chosen such that (14.14) holds. The little group of π, denoted by $\tilde{\mathbb{E}}(2)$, is different. It consists of all $A \in SL(2,\mathbb{C})$ satisfying $A\underline{\pi}A^\dagger = \underline{\pi}$, whence A is of the form

$$A = \begin{pmatrix} \mathrm{e}^{\mathrm{i}\varphi/2} & a\mathrm{e}^{-\mathrm{i}\varphi/2} \\ 0 & \mathrm{e}^{-\mathrm{i}\varphi/2} \end{pmatrix}, \tag{14.24}$$

with $a \in \mathbb{C}$. This group is a 2 : 1 covering of the group of Euclidean motions $\mathbb{E}(2)$ in two dimensions. Indeed, an element of $\tilde{E}(2)$ is characterized by a pair $(a, \mathrm{e}^{\mathrm{i}\varphi/2}$, and if we associate to this the Euclidean motion (Re a, Im a; R_φ), consisting of the translation (Re a, Im a) and the rotation R_φ by the angle φ, we obtain a homomorphism with kernel $(0, 0; \pm 1)$. Hence,

$$\frac{\tilde{\mathbb{E}}(2)}{(0,0;\pm 1)} \cong \mathbb{E}(2)\,. \tag{14.25}$$

Next, we have to determine the irreducible unitary representations of $\tilde{\mathbb{E}}(2)$. This is done along the same lines as for $\tilde{\mathcal{P}}^0$. First we choose improper eigenstates for the *translations.* We then have two cases. Either the *momenta* lie on a circle with radius $\rho > 0$ or the orbit in $\mathbb{R}^2$ under $U(1)$ consists only of the point $\mathbf{0}$. In the first case the representations of $\tilde{\mathbb{E}}(2)$ are infinite dimensional. Since this means that there are infinitely many degrees of freedom (continuous spin) these massless representations appear to be unphysical. Therefore, we consider here only the second case, where the two-dimensional *translations* are represented trivially. Then the little group is $U(1)$. Its irreducible unitary representations are one-dimensional

$$\vartheta^{(\lambda)} : \mathrm{e}^{\mathrm{i}\varphi/2} \mapsto \mathrm{e}^{\mathrm{i}\lambda\varphi};\ \ \lambda = 0, \pm 1/2, \pm 1, \ldots$$

Thus, the degeneracy parameter λ takes only a *single value* in an irreducible representation for $m = 0$, and the action of $\tilde{\mathbb{E}}(2)$ on $|\pi, \lambda\rangle$ is given by

$$U(a, e^{i\varphi/2})|\pi, \lambda\rangle = e^{i\lambda\varphi}|\pi, \lambda\rangle \ . \tag{14.26}$$

The formulae in (14.18) remain valid for $m = 0$ if $D^{(s)}$ of $SU(2)$ is replaced by $\vartheta^{(\lambda)}$ of $U(1)$. The Wigner *rotation* is now an element of $\tilde{\mathbb{E}}(2)$.

The boosts $L(p)$ can again be chosen such that $|p, \lambda\rangle$ describe helicity states.

14.3 On Mackey's Theory of Induced Representations

We consider the following situation. Let G be a locally compact group. (All topological spaces are assumed to satisfy the second axiom of countability.) Let H be a closed subgroup of G and consider the homogeneous space $X = G/H$, the space of all left cosets gH, $g \in G$. $\pi : G \to X$ denotes the canonical mapping, defined by $\pi(g) = gH$. X is a transitive G-space with the action

$$gx = \pi(gs) \ , \quad g, s \in G, \ x = \pi(s) \ .$$

We equip X with the quotient topology. Below we shall use the fact that there is a continuous section $\sigma : X \to G$, which satisfies per definition $\pi \circ \sigma = id$. We also use the fact that G has a left invariant Haar measure on the σ-algebra of Borel sets, which is unique, up to a normalization factor. On X one can easily construct quasi-invariant measures, which means that null sets are transformed under the action of G into null sets. These are all mutually absolutely continuous. If μ is such a measure and $\mu^g(E) := \mu(g^{-1}E)$, then μ and μ^g are equivalent and $d\mu^g = (d\mu^g / d\mu)\, d\mu$, where $d\mu^g / d\mu$ is the Radon–Nikodym derivative, which we will denote by $\rho_g(x)$. This Borel function satisfies

$$\rho_{g_1 g_2}(x) = \rho_{g_1}(x)\rho_{g_2}\left(g_1^{-1}x\right) \ . \tag{14.27}$$

Let now $L : H \to \mathcal{U}(\mathcal{H})$ be a unitary representation of H in the Hilbert space $\mathcal{H}$ ($\mathcal{U}(\mathcal{H})$ denotes the unitary operators of $\mathcal{H}$). Consider maps $f : G \to \mathcal{H}$ such that:

1. $(\Phi, f(g))$ is measurable for all $\Phi \in \mathcal{H}$;
2. $f(gh) = L(h^{-1})f(g)$, $h \in H$;
3. $\int_{G/H} \|f\|^2 \, d\mu < \infty$.

For the last condition note that $\|f\|$ depends only on equivalence classes gH. These functions form a Hilbert space with respect to the scalar product

$$(f_1, f_2) = \int\limits_{G/H} \langle f_1, f_2 \rangle_{\mathcal{H}} \, d\mu \ . \tag{14.28}$$

The *induced representation* of G in this Hilbert space is defined by

$$\left(U_g^L f\right)(s) = \sqrt{\rho_g(\pi(s))}\, f(g^{-1}s) \ . \tag{14.29}$$

One easily verifies that this is indeed a representation that is unitary.

Reformulation 1

We choose a section σ as described above, and define $\psi(x) = f(\sigma(x))$ (Fig. 14.2).

Figure 14.2 f can be recovered from ψ

$$f(g) = f(\sigma(x) \underbrace{\sigma(x)^{-1} g}_{\in H}) = L(g^{-1}\sigma(x))\psi(x) \ ,$$
$$x = \pi(g) \ . \tag{14.30}$$

We now rewrite (14.29) in terms of ψ (for simplicity we assume $\rho_g(s) = 1$). Because of the last equation it is natural to define the transformation of ψ by

$$\left(U_g^L f\right)(s) =: L(s^{-1}\sigma(x))(V_g^L \psi)(x) \ , \quad x = \pi(s) \ .$$

Here, the left-hand side is

$$\begin{aligned} f(g^{-1}s) &= L(s^{-1}g\sigma(\pi(g^{-1}s)))\psi(\underbrace{\pi(g^{-1}s)}_{g^{-1}x}) \\ &= L(s^{-1}\sigma(x))L(\sigma(x)^{-1}g\sigma(g^{-1}x))\psi \\ &\quad \times (g^{-1}x) \ . \end{aligned}$$

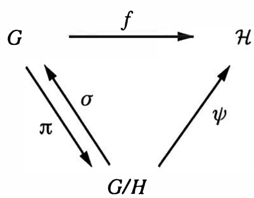


Fig. 14.2 Diagram for the Map ψ

Hence we obtain, including the case of a nontrivial ρ_g,

$$(V_g^L\psi)(x) = \sqrt{\rho_g(x)}L(\underbrace{\sigma(x)^{-1}g\sigma(g^{-1}x))}_{\in H} \times \psi(g^{-1}x) . \quad (14.31)$$

This is a unitary representation in the Hilbert space $L^2(G/H, \mu; \mathcal{H})$ of $\mathcal{H}$-valued functions. (Verify the representation property.)

Reformulation 2

Embed L into a representation $\tilde{L}$ of G; $\tilde{L}$ need not be unitary. So we assume that there is a Hilbert space $\tilde{\mathcal{H}}$ and a representation $\tilde{L}$ of G in $\tilde{\mathcal{H}}$, such that $\mathcal{H}$ can be identified with a Hilbert subspace of $\tilde{\mathcal{H}}$ and $\tilde{L}(h)u = L(h)u$ for all $h \in H, u \in \mathcal{H}$. We associate to each $f : G \to \mathcal{H}$, satisfying the properties 1–3 above, the map $\varphi : G \to \tilde{\mathcal{H}}$, defined by

$$\varphi(g) = \tilde{L}(g)f(g) . \quad (14.32)$$

The covariance condition 2 then becomes $\varphi(gh) = \varphi(g)$ for all $h \in H$, i.e., φ depends only on the coset $[g] \in G/H$. So φ induces the map $\omega : X = G/H \to \tilde{\mathcal{H}}$,

$$\omega(x) = \varphi(g) , \quad x = [g] = \pi(g) . \quad (14.33)$$

For φ the transformation law becomes $(U_g\varphi)(s) = \tilde{L}(g)\varphi(g^{-1}s)$. This induces

$$(U_g\omega)(x) = \tilde{L}(g)\omega(g^{-1}x) . \quad (14.34)$$

In the space of maps $\omega : X \to \tilde{\mathcal{H}}$ we introduce a scalar product, such that the transformation (14.34) is unitary. For this consider for $x \in X$ a group element $g \in G$ with $gx_0 = x$, where $x_0 = [e] = H$, and define the subspace

$$\mathcal{H}_x = \tilde{L}(g)(\mathcal{H}) . \quad (14.35)$$

This depends only on $[g]$. In $\mathcal{H}_x$ define the scalar product

$$\langle u, v\rangle_x = \langle \tilde{L}(g^{-1})u, \tilde{L}(g^{-1})v\rangle_{\mathcal{H}} . \quad (14.36)$$

This is well defined since $L(h)$ is unitary. Note also that

$$\mathcal{H}_{sx} = \tilde{L}(s)(\mathcal{H}_x) , \quad s \in G , \quad (14.37)$$

and

$$\langle \tilde{L}(s)u, \tilde{L}(s)v\rangle_{sx} = \langle u, v\rangle_x . \quad (14.38)$$

The map ω satisfies $\omega(x) \in \mathcal{H}_x$. The scalar product of two such maps ω_1, ω_2 is defined by

$$(\omega_1, \omega_2) = \int_X \langle \omega_1(x), \omega_2(x)\rangle_x \, d\mu(x) . \quad (14.39)$$

From now on we consider ω's in the corresponding Hilbert space $\mathcal{H}_\omega$, and we assume that the measure μ is invariant.

Representation (14.34) is unitary in $\mathcal{H}_\omega$. Indeed, using (14.38) we have

$$\langle (U_s\omega_1)(x), ((U_s\omega_2)(x)\rangle_x = \langle \omega_1(s^{-1}x) , \omega_2(s^{-1}x)\rangle_{s^{-1}x} .$$

Together with the invariance of μ on G/H the claim follows.

Remarks 14.1

1. The scalar product (14.38) is more complicated than that for the original maps f. This is the price we have to pay for the simple transformation law (14.34) for $\omega \in \mathcal{H}_\omega$.
2. Representation $\tilde{L}|H$ is typically *not* irreducible. To arrive at irreducible representations of H we have to impose subsidiary conditions. This will become important in Sect. 14.A when we discuss free fields for arbitrary spin.
3. There is also a description in terms of *G-Hilbert space bundles* [14.9], which is completely equivalent to what we have done.

14.3.1 Application to Semidirect Products

We now specialize the theory of induced representations to semidirect products $G = A \rtimes H$ relative to an action of H on A, $a \mapsto ha$. (Examples: The inhomogeneous linear groups and certain subgroups, for instance the inhomogeneous Lorentz group.) Both groups are assumed to be locally compact, and we will only consider the case when A is Abelian. For this class Mackey's theory guarantees that the induction process provides all irreducible unitary representations.

We note that A and H can be regarded as subgroups of G, A being a closed normal subgroup. Furthermore, $G = AH$, $A \cap H = e$, $ha = hah^{-1}$. This can be regarded as an internal characterization of semidirect products.

Let $\hat{A}$ be the character group of A, i. e., the set of continuous homomorphisms of A into the group of complex numbers of modulus 1. Under pointwise multiplication this set becomes a group. Relative to the topology of uniform convergence on compacta it is locally compact and satisfies the second axiom of countability. For $x \in \hat{A}$ we denote its value on $a \in A$ by $\langle x, a\rangle$. The action of H on A induces an action of H on $\hat{A}$ by $\langle hx, a\rangle = \langle x, h^{-1}a\rangle$; $x \mapsto hx$ is well defined and continuous. We choose a point $x_0 \in \hat{A}$ and denote by $Hx_0 = X$ the orbit of x_0 in $\hat{A}$. Let H_0 be the stabilizer of H at x_0, i. e., $H_0 = \{h : h \in H,\ hx_0 = x_0\}$. We extend the action of H on $\hat{A}$ to one by all of G, assuming that A acts trivially. Note that if $\alpha(g)$ denotes the inner automorphism on A, $\alpha(g)(a) = gag^{-1}$, then the extended action is given by $\langle gx, a\rangle = \langle x, \alpha(g)^{-1}(a)\rangle$. This turns X into a G-space. The stability subgroup of G is $G_0 = A \rtimes H_0$.

For what follows we note that the map $G/G_0 \to X$, $[g] \mapsto [g]x_0$ (defined with representatives) is a G-isomorphism (verify this). Note that obviously $G/G_0 \cong H/H_0$, so we can also identify X with H/H_0.

Let $D(h)$ be a unitary representation of H_0 in the Hilbert space $\mathcal{H}$ and consider the extension $L(ah) = \langle x_0, a\rangle D(h)$ to G_0. For this situation we can use the transformation law (14.31). Thanks to the G-isomorphism just mentioned, we can regard the functions ψ in (14.31) as functions on X. With this reinterpretation we have to use instead of the sections $\sigma : G/G_0 \to G$ maps $c : X \to H \subset G$ with $c(x)x_0 = x$, in terms of which (14.31) becomes for $\rho_g \equiv 1$

$$(V_g\psi)(x) = L(c(x)^{-1}gc(g^{-1}x))\psi(g^{-1}x) . \quad (14.40)$$

For $g = a \in A$ this gives

$$\begin{aligned}(V_a\psi)(x) &= \langle x_0, c(x)^{-1}ac(x)\rangle\psi(x) \\ &= \langle c(x)x_0, a\rangle\psi(x) = \langle x, a\rangle\psi(x) ,\end{aligned}$$

and for $g = h \in H$ we obtain

$$(V_h\psi)(x) = D(c(x)^{-1}hc(h^{-1}x))\psi(h^{-1}x) .$$

Since $V_{ah} = V_a V_h$ we obtain the unitary representation

$$(V_{ah}\psi)(x) = \langle x, a\rangle D(c(x)^{-1}hc(h^{-1}x))\psi(h^{-1}x) , \quad (14.41)$$

of $G = A \rtimes H$ in the Hilbert space $L^2(X, \mu; \mathcal{H})$, where μ now denotes the transported measure to X (assumed to be invariant).

Mackey's theory establishes the following important result (for detailed proofs, see [14.10]):

Theorem 14.3 Mackey

Let us choose, for each H-orbit Ω in $\hat{A}$, a point x_Ω on Ω, and an irreducible representation D of the stability subgroup H_Ω at the point x_Ω. Then the representation $V^{D,\Omega}$, given by (14.41), is irreducible. Two such representations are equivalent if and only if the orbits coincide, and the representations of the stabilizer are equivalent. If the H-orbit structure of $\hat{A}$ satisfies a certain smoothness property, then each irreducible representation is equivalent to some $V^{D,\Omega}$.

In the Appendix we indicate Mackey's strategy.

Let us specialize this important result for the universal covering group $\mathbb{R}^4 \rtimes SL(2,\mathbb{C})$ of $\mathcal{P}_+^\uparrow$. With the notation introduced in Sect. 3, Eq. (14.41) becomes, for example, for the orbit H_m^+

$$\begin{aligned}(U(a,A)f)(p) &= \mathrm{e}^{ipa} D\left(L(p)^{-1}AL\left(\Lambda_A^{-1}p\right)\right) \\ &\quad \times f(\Lambda_A^{-1}p) , \\ f &\in L^2\left(H_m^+, \mathrm{d}\Omega_m; \mathcal{H}\right) .\end{aligned} \quad (14.42)$$

For $D = D^{(s)}$ this agrees with (14.19). For the applications in the next section we introduce a construction similar to the reformulation 2 above.

Let us assume that the Hilbert space $\mathcal{H}$ is a subspace of a Hilbert space $\tilde{\mathcal{H}}$, and $\tilde{D}$ is a representation of $SL(2,\mathbb{C})$ in $\tilde{\mathcal{H}}$, not necessarily unitary, such that the restriction of $\tilde{D}$ to $SU(2)$ in $\mathcal{H}$ is equal to D. (The restriction may, however, be reducible in $\tilde{\mathcal{H}}$.) Let $\mathcal{H}_p = \tilde{D}(L(p))(\mathcal{H})$, with the inner product

$$\langle u, v\rangle_p = \langle \tilde{D}(L(p)^{-1})u, \tilde{D}(L(p)^{-1})v\rangle_{\mathcal{H}} . \quad (14.43)$$

Consider Borel maps $\psi : H_m^+ \to \tilde{\mathcal{H}}$ with $\psi(p) \in \mathcal{H}_p$. Clearly, if

$$\psi(p) := \tilde{D}(L(p))f(p) , \quad (14.44)$$

then

$$\langle \psi_1(p), \psi_2(p)\rangle_p = \langle f_1(p), f_2(p)\rangle_{\mathcal{H}} . \quad (14.45)$$

In terms of ψ (14.42) becomes (abusing notation)

$$(U(a,A)\psi)(p) = \mathrm{e}^{ipa}\tilde{D}(A)\psi\left(\Lambda_A^{-1}p\right) . \quad (14.46)$$

We choose ψ in the Hilbert space of maps with finite norm belonging to the scalar product

$$(\psi_1, \psi_2) = \int_{H_m^+} \langle \psi_1(p), \psi_2(p)\rangle_p \mathrm{d}\Omega_m(p) . \quad (14.47)$$

This construction gives a unitary representation of $\tilde{\mathcal{P}}_+^\uparrow$ which is *not irreducible* when $\tilde{D}|SU(2)$ is reducible in $\tilde{\mathcal{H}}$. In order to obtain irreducible representations, we have to impose *subsidiary conditions*. This brings us to the next topic.

14.4 Free Classical and Quantum Fields for Arbitrary Spin, Spin, and Statistics

With the developed group theoretical tools we can now give an elegant approach to fields with arbitrary spin (see also [14.11]). We first consider classical fields.

14.4.1 Classical Fields for Arbitrary Spin and Positive Mass

A classical relativistic field $\psi_\alpha(x)$ is a solution of a system of Lorentz invariant field equations. Under $\tilde{\mathcal{P}}_+^\uparrow \equiv \tilde{\mathcal{P}}^0$ the field transforms according to

$$\psi'_\alpha(x') = S(A)_{\alpha\beta}\psi_\beta(x) , \quad x' = \Lambda_A x + a . \quad (14.48)$$

Here, $A \mapsto S(A)$ is a finite-dimensional representation of $SL(2,\mathbb{C})$. We consider only free fields. Then the solution space is linear and hence we can define a representation of $\tilde{\mathcal{P}}^0$ by

$$(U(a,A)\psi)_\alpha(x) = S(A)_{\alpha\beta}\psi_\beta\left(\Lambda_A^{-1}(x-a)\right) . \quad (14.49)$$

In this section, we construct systems of linear field equations, such that the positive frequency solutions give rise to an irreducible unitary Wigner representation $(m,s), m > 0$.

$2s+1$ Component Field Equation

For the extension of $D^{(s)}$ to $SL(2,\mathbb{C})$ we choose, in standard notation, the representation $D^{(s,0)}$ that we also denote by $D^{(s)}$. Then (14.44) becomes

$$\varphi_\alpha(p) = \sum_{\lambda=-s}^{s} D^{(s)}_{\alpha\lambda}(L(p))f_\lambda(p) , \quad (14.50)$$

and the norm belonging to (14.47) is

$$\|\varphi\|^2 = \int \varphi^\dagger(p)D^{(s)}\left(\frac{\hat{p}}{m}\right)\varphi(p)\,\mathrm{d}\Omega_m(p) . \quad (14.51)$$

The 'hat' symbol on a 2×2 matrix A is defined by $\hat{A} = \varepsilon\bar{A}\varepsilon^{-1}$ where ε is the standard symplectic matrix. For $A \in SL(2,\mathbb{C})$ one easily finds $\hat{A} = (A^\dagger)^{-1}$. In (14.51) we have used (14.13). Transformation (14.46) becomes

$$(U(a,A)\varphi(p) = \mathrm{e}^{\mathrm{i}pa}D^{(s)}(A)\varphi(\Lambda_A^{-1}p) . \quad (14.52)$$

This is precisely of the form (14.49) in momentum space, with $S(A) = D^{(s,0)}(A)$. Since the restriction of $D^{(s,0)}$ to $SU(2)$ is $D^{(s)}$, representation (14.52) is irreducible and equivalent to the Wigner representation (m,s). No subsidiary conditions have to be imposed. If we pass to x-space by

$$\varphi_\alpha(x) = (2\pi)^{-3/2}\int \varphi_\alpha(p)\mathrm{e}^{-\mathrm{i}px}\mathrm{d}\Omega_m(p) , \quad (14.53)$$

then $\varphi_\alpha(x)$ satisfies only the Klein–Gordon equation

$$\left(\Box + m^2\right)\varphi_\alpha(x) = 0 , \quad \alpha = -s,\ldots,+s . \quad (14.54)$$

Beside the positive frequency solutions, this equation has also negative frequency solutions, which span an irreducible unitary representation belonging to the orbit H_m^- and spin s.

$2(2s+1)$ Component Field Equation

Instead of the extension $D^{(s,0)}$ we could have used $D^{(0,s)}$. This is equivalent to the representation $\hat{D}^{(s)}(A) := D^{(s,0)}(\hat{A}) = D^{(s,0)}(A)^{\dagger-1}$. For this case we introduce the *spinor amplitudes*

$$\chi^{\dot{\alpha}}(p) = \sum_{\lambda=-s}^{s} \hat{D}^{(s)}_{\dot{\alpha}\lambda}(L(p))f_\lambda(p) . \quad (14.55)$$

The scalar product now becomes

$$(\chi_1,\chi_2) = \int \chi_1^\dagger(p)D^{(s)}\left(\frac{p}{m}\right)\chi_2(p)\,\mathrm{d}\Omega_m(p) . \quad (14.56)$$

Part B | 14.4

The χ-fields transform according to

$$U(a,A)\chi(p) = e^{ipa}\hat{D}^{(s)}(A)\chi\left(\Lambda_A^{-1}p\right) . \tag{14.57}$$

In this case $S(A)$ in (14.49) is $\hat{D}^{(s)}$.

The fields φ and χ are, of course, not independent. We claim that

$$\chi(p) = D^{(s)}\left(\frac{\hat{\underline{p}}}{m}\right)\varphi(p) ,$$
$$\varphi(p) = D^{(s)}\left(\frac{\underline{p}}{m}\right)\chi(p) . \tag{14.58}$$

For instance,

$$\begin{aligned}\chi(p) &= \hat{D}^{(s)}(L(p))f(p)\\ &= \hat{D}^{(s)}(L(p))D^{(s)}(L(p))^{-1}\varphi(p)\\ &= D^{(s)}(\hat{L}(p)L(p)^{-1})\varphi(p)\\ &= D^{(s)}\left(\frac{\hat{\underline{p}}}{m}\right)\varphi(p) .\end{aligned}$$

Equations (14.58) are the generalizations of the Dirac equation for $s = 1/2$

$$\hat{\underline{p}}\varphi(p) = m\chi(p) ,$$
$$\underline{p}\chi(p) = m\varphi(p) . \tag{14.59}$$

Imposing (14.58) as subsidiary equations provides again an irreducible representation in the space of $2 \times (2s+1)$-component fields

$$\psi(p) = \begin{pmatrix}\varphi(p)\\ \chi(p)\end{pmatrix} , \tag{14.60}$$

transforming according to the reducible representation

$$S(A) = \begin{pmatrix}D^{(s)}(A)\\ \hat{D}^{(s)}(A)\end{pmatrix} . \tag{14.61}$$

In x-space Eqs. (14.58) become

$$D^{(s)}(i\hat{\underline{\partial}})\varphi(x) = m^{2s}\chi(x) ,$$
$$D^{(s)}(i\underline{\partial})\chi(x) = m^{2s}\varphi(x) . \tag{14.62}$$

In addition, ψ satisfies, of course, the Klein–Gordon equation.

We also introduce generalizations of the Dirac matrices. Since $D^{(s)}(\underline{p})$ is a homogeneous polynomial of degree $2s$ in p, we can set

$$D^{(s)}(\underline{p}) = \sigma^{\mu_1\ldots\mu_{2s}}p_{\mu_1}\cdots p_{\mu_{2s}} ,$$
$$D^{(s)}(\hat{\underline{p}}) = \hat{\sigma}^{\mu_1\ldots\mu_{2s}}p_{\mu_1}\cdots p_{\mu_{2s}} . \tag{14.63}$$

The generalized Dirac matrices are defined by

$$\gamma^{\mu_1\ldots\mu_{2s}} = \begin{pmatrix}0 & \sigma^{\mu_1\ldots\mu_{2s}}\\ \hat{\sigma}^{\mu_1\ldots\mu_{2s}} & 0\end{pmatrix} . \tag{14.64}$$

With these we can write the field equations (14.62) as

$$\left[(-i)^{2s}\gamma^{\mu_1\ldots\mu_{2s}}\partial_{\mu_1}\cdots\partial_{\mu_{2s}} + m^{2s}\right]\psi(x) = 0 . \tag{14.65}$$

For $s = 1/2$ this reduces to the Dirac equation. Fields of this type have been considered, for instance, in [14.12, 13].

Bargmann–Wigner Fields

These fields are constructed with yet another extension of $D^{(s)}$ to $SL(2,\mathbb{C})$. We realize the Wigner representation (m, s) in the Hilbert space

$$\mathcal{H}^{(m,s)} = \left\{ f_{\lambda_1\ldots\lambda_{2s}}(p) \middle| \sum_{(\lambda)}\int |f_{\lambda_1\ldots\lambda_{2s}}(p)|^2 \times d\Omega_m(p) < \infty \right\} , \tag{14.66}$$

where the functions f are symmetric in the two-valued indices. So the functions f are maps from H_m^+ into the $2s$-fold symmetric tensor product of $\mathbb{C}^2$. The Wigner representation is

$$\begin{aligned}&(U^{(m,s)}(a,A)f)_{\lambda_1\ldots\lambda_{2s}}(p)\\ &= e^{ipa}\sum_{(\lambda)}\prod_j (R(p,A))_{\lambda_j\lambda_j'}f_{\lambda_1'\ldots\lambda_{2s}'}\left(\Lambda_A^{-1}p\right) .\end{aligned} \tag{14.67}$$

Now, we define generalized Dirac spinors. Let

$$B_{a\lambda}(p) = \begin{pmatrix}L_{\alpha\lambda}(p)\\ \hat{L}_{\dot{\alpha}\lambda}(p),\end{pmatrix} , \quad a = (\alpha, \dot{\alpha}) , \tag{14.68}$$

and define

$$\psi_{a_1\ldots a_{2s}}(p) = \sum_{(\lambda)} \prod_j B_{a_j\lambda_j}(p) f_{\lambda_1\ldots\lambda_{2s}}(p) . \quad (14.69)$$

Dropping indices, we also write

$$\psi(p) = \left(\bigotimes_j B_j(p) \right) f(p) .$$

The scalar product (14.43) becomes, using (14.59),

$$\langle \psi_1(p), \psi_2(p) \rangle_p = \frac{1}{2} \psi_1^\dagger(p) \bigotimes_j \gamma^0_{(j)} \psi_2^\dagger , \quad (14.70)$$

where

$$\gamma^\mu_{(j)} = \mathbf{1} \otimes \cdots \otimes \gamma^\mu \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1}$$

($2s$ factors, γ^μ at position j). As a result of the identity,

$$\frac{1}{m} \gamma^\mu p_\mu \begin{pmatrix} L_{\alpha\lambda}(p) \\ \hat{L}_{\dot{\alpha}\lambda}(p), \end{pmatrix} = \begin{pmatrix} L_{\alpha\lambda}(p) \\ \hat{L}_{\dot{\alpha}\lambda}(p) \end{pmatrix} ,$$

the $\psi(p)$ satisfy the *Bargmann–Wigner equations*

$$\left(\gamma^\mu_{(j)} p_\mu - m \right) \psi = 0 . \quad (14.71)$$

There are, by construction, no other subsidiary conditions (show this).

For the transformation law of the Bargmann–Wigner fields one readily finds

$$(U(a, A)\psi(p) = e^{ipa} \left(\bigotimes_j S_j(A) \right) \psi \left(\Lambda_A^{-1} p \right) , \quad (14.72)$$

where each $S_j(A)$ is equal to the reducible Dirac representation $D^{(1/2)} \bigoplus \hat{D}^{(1/2)}$:

$$S(A) = \begin{pmatrix} A & 0 \\ 0 & \hat{A} \end{pmatrix} .$$

This shows that $\psi_{a_1\ldots a_{2s}}$ is a symmetric *multi-Dirac spinor*.

Pauli–Fierz Fields

Let m, n be two integers ≥ 0 with $m + n = 2s$. The Pauli–Fierz spinor fields are defined by

$$\phi^{\dot{\beta}_1\ldots\dot{\beta}_m}_{\alpha_1\ldots\alpha_n}(p) = \prod_{j=1}^{n} L_{\alpha_j\lambda_j}(p) \times \prod_{k=1}^{m} \hat{L}_{\dot{\beta}_k\lambda_k}(p) f_{\lambda_1\ldots\lambda_n;\mu_1\ldots\mu_m}(p) , \quad (14.73)$$

where f is separately symmetric in the indices λ and μ.

The identities

$$\underline{p}\hat{L}(p) = mL(p) ,$$
$$\hat{\underline{p}}L(p) = m\hat{L}(p) . \quad (14.74)$$

imply the *Pauli–Fierz equations* [14.14]

$$p^{\alpha\dot{\beta}} \phi^{\dot{\beta}_2\ldots\dot{\beta}_m}_{\alpha\alpha_1\ldots\alpha_n} = m\phi^{\dot{\beta}\dot{\beta}_2\ldots\dot{\beta}_m}_{\alpha_1\ldots\alpha_n} ,$$
$$p_{\alpha\dot{\beta}} \phi^{\dot{\beta}\dot{\beta}_2\ldots\dot{\beta}_m}_{\alpha_1\ldots\alpha_n} = m\phi^{\dot{\beta}_2\ldots\dot{\beta}_m}_{\alpha\alpha_1\ldots\alpha_n} . \quad (14.75)$$

Different choices of m, n lead to different fields. As long as we do not consider reflections or interactions, all these fields are, by construction, equivalent.

Rarita–Schwinger Fields

For practical calculations with half-integer spin $\geq 3/2$, fields introduced by Rarita and Schwinger are very useful. One can arrive at these starting from the Pauli–Fierz fields. For details, I refer to [14.7]. If $s = 3/2$, the Rarita–Schwinger field has a Dirac and a vector index; notation: $\psi_\mu(x)$, where the Dirac index is not written. From the construction one obtains the *Rarita–Schwinger equations*

$$\left(\gamma^\nu p_\nu - m \right) \psi_\mu = 0 , \quad (14.76)$$

plus the subsidiary condition

$$\gamma^\mu \psi_\mu = 0 . \quad (14.77)$$

14.4.2 Free Quantum Fields, Spin Statistics

So far we have only considered one-particle states, transforming irreducibly under $\tilde{\mathcal{P}}^0$ (elementary systems in the sense of Wigner). It should be said at this point that from the transformation law alone we do not know whether the system is elementary or composite in the

usual sense, in which an electron is *elementary* and a deuteron is composite. (For an interesting dispute on this delicate issue between Heisenberg and Wigner, see the discussion after Heisenberg's talk at the Dirac conference [14.15].)

In a theory of fundamental interactions, like the Standard model of particle physics, the elementary systems in the sense of Wigner, span a proper subspace $\mathcal{H}_1 \subset \mathcal{H}$ that is invariant under the representation $U(a, A)$ of $\tilde{\mathcal{P}}^0$ in the total space $\mathcal{H}$.

We discuss here only the Hilbert space of an arbitrary number of noninteracting particles. This is essential for the formulation of the scattering problem (description of asymptotic states).

Fock Space over (m, s)

Let $\mathcal{F}_1$ be the one-particle space $L^2(H_m^+, \mathrm{d}\Omega_m; \mathbb{C}^{2s+1})$ carrying the Wigner representation (m, s)

$$(U_1(a,A)f)(p) = \mathrm{e}^{\mathrm{i}pa} D^{(s)}(L(p)^{-1} \times AL(\Lambda_A^{-1}p)) f(\Lambda_A^{-1}p) \,. \quad (14.78)$$

The space of N-particle states is

$$\mathcal{F}_N = \mathcal{F}_1 \otimes_{s,a} \cdots \otimes_{s,a} \mathcal{F}_1 \quad (N \text{ times}) \,, \quad (14.79)$$

where $\otimes_{s,a}$ denotes the symmetric or antisymmetric tensor product. Explicitly,

$$\mathcal{F}_N = \Big\{ f(p_1, \lambda_1, \ldots p_N, \lambda_N) \Big| \, f \ \begin{matrix}\text{symmetric or}\\ \text{antisymmetric}\end{matrix} \,, \; \|f\|_N^2 < \infty \Big\} \,,$$

with

$$\|f\|_N^2 = \sum_{(\lambda)} \int |f(p_1, \lambda_1, \ldots p_N, \lambda_N)|^2 \times \mathrm{d}^N \Omega_m(p) \,.$$

The Fock space is the direct Hilbert sum ($\mathcal{F}_0 := \mathbb{C}$)

$$\mathcal{F} = \bigoplus_{N=0}^{\infty} \mathcal{F}_N \,. \quad (14.80)$$

An element $f \in \mathcal{F}$ is a sequence $f = (f^{(0)}, f^{(1)}, \ldots)$, with

$$\|f\|^2 = \sum_{N=0}^{\infty} \|f^{(N)}\|_N^2 \,.$$

The special state $\Omega_F = (1, 0, \ldots)$ is the *Fock vacuum.* The representation U_1 in $\mathcal{F}_1$ induces in a natural manner representations U_N in $\mathcal{F}_N$ and U in $\mathcal{F}$. (On $\mathcal{F}_0$ the representation is trivial: invariance of the Fock vacuum.)

Interpretation: Let $f \in \mathcal{F}$, $f = \{f^{(N)}\}$, then $|f^{(N)}(p_1, \lambda_1, \ldots, p_N, \lambda_N)|^2 \, \mathrm{d}^N \Omega_m(p)$ is the probability measure in momentum space for given spin components $\lambda_1, \ldots, \lambda_N$.

In what follows, $\mathcal{F}_\infty$ denotes the subspace of $\mathcal{F}$, whose elements have only a finite number of nonvanishing components. On $\mathcal{F}_\infty$ one can introduce the standard creation and annihilation operators $a(g), a^\dagger(g)$ for $g \in \mathcal{F}_1$. For instance, if $f \in \mathcal{F}_\infty$, then

$$(a(g)f)^{(n-1)}(p_1, \lambda_1, \ldots, p_n, \lambda_n) = \sqrt{n} \int \mathrm{d}\Omega_m(p) \sum_\lambda g * (p, \lambda) f^{(n)} \times (p, \lambda, p_1, \lambda_1, .., p_{n-1}, \lambda_{n-1}) \,.$$

On $\mathcal{F}_\infty$ the creation and annihilation operators are adjoint to each other and satisfy

1. $[a(g_1), a^\dagger(g_2)]_\pm = (g_1, g_2)_1$ ($\pm$ for symmetric (antisymmetric) tensor products);
2. $U(a,A) a^\dagger(g) U^{-1}(a,A) = a^\dagger(g_{(a,A)})$, $g_{(a,A)} = U_1(a,A) g$.

(Subtleties connected with unbounded operators are treated in [14.16].)

$2s+1$ Component Quantum Fields

Now, we introduce quantum versions of the fields constructed in Sect. 5.1.1. Let $\{f_k(p, \lambda)\}$ be an orthonormal basis in $\mathcal{F}_1$, and

$$u_\alpha^{(k)}(x) = (2\pi)^{-3/2} \int \mathrm{d}\Omega_m(p) \times \sum_\lambda D_{\alpha\lambda}^{(s)}(L(p)) f_k(p, \lambda) \mathrm{e}^{-\mathrm{i}px} \,,$$

$$v_\alpha^{(k)}(x) = (2\pi)^{-3/2} \int \mathrm{d}\Omega_m(p) \times \sum_\lambda D_{\alpha\lambda}^{(s)}(L(p)\varepsilon) f_k^*(p, \lambda) \mathrm{e}^{\mathrm{i}px} \,. \quad (14.81)$$

With this we define the quantum field (operator-valued distribution)

$$\varphi_\alpha(x) = \sum_k \Big[a(f_k) u_\alpha^{(k)}(x) + a^\dagger(f_k) v_\alpha^{(k)}(x) \Big] \,. \quad (14.82)$$

This expression becomes more transparent if we write symbolically

$$a^\dagger(f) = \int \mathrm{d}\Omega_m(p) \sum_\lambda a(p,\lambda) f(p,\lambda) . \qquad (14.83)$$

Then we get

$$\varphi_\alpha(x) = \frac{1}{(2\pi)^{3/2}} \int \mathrm{d}\Omega_m(p) \sum_\lambda \times \Big\{ D^{(s)}_{\alpha\lambda}(L(p)) a(p,\lambda) \mathrm{e}^{-\mathrm{i}px} + D^{(s)}_{\alpha\lambda}(L(p)\varepsilon) a^*(p,\lambda) \mathrm{e}^{\mathrm{i}px} \Big\} . \qquad (14.84)$$

Remarks 14.2

1. We have only introduced one sort of particles. The generalization to the case, where the antiparticles are different, is obvious.
2. The factor of $a(p,\lambda)$, namely $D^{(s)}_{\alpha\lambda}(L(p))\mathrm{e}^{-\mathrm{i}px} \equiv u_\alpha(x,\lambda)$ is a plane-wave positive frequency solution of the classical field in Sect. 5.1.1. This factor and the corresponding one for $a^\dagger(p,\lambda)$ are chosen such that $\varphi_\alpha(x)$ transforms as

$$U(a,A)\varphi_\alpha(x)U^{-1}(a,A) = \sum_\beta D^{(s)}_{\alpha\beta}(A^{-1})\varphi_\beta(\Lambda_A x + a) . \qquad (14.85)$$

The verification of this is straightforward.

Now we come to a crucial point. We shall see that the field is *only local if we choose the standard connection between spin and statistics*. For this we compute $[\varphi_\alpha(x), \varphi^\dagger_\beta(y)]_\pm$, using

$$[a(p,\lambda), a^\dagger(p',\lambda') = \delta_{\lambda'\lambda} 2p^0 \delta^{(3)}(\mathbf{p}' - \mathbf{p}) . \qquad (14.86)$$

(We proceed formally, but the derivation can easily be rewritten in a mathematically rigorous manner.) A short calculation, using (14.13), leads to the important result

$$\Big[\varphi_\alpha(x), \varphi^\dagger_\beta(y)\Big]_\pm = \frac{1}{(2\pi)^3} \int \mathrm{d}\Omega_m(p) D^{(s)}_{\alpha\beta}\left(\frac{p}{m}\right) \times \Big[\mathrm{e}^{-\mathrm{i}p(x-y)} \pm (-1)^{2s} \mathrm{e}^{\mathrm{i}p(x-y)}\Big] . \qquad (14.87)$$

If and only if $\pm(-1)^{2s} = -1$, that is if the normal connection between spin and statistics holds, we get a local field

$$\Big[\varphi_\alpha(x), \varphi^\dagger_\beta(y)\Big]_\pm = \mathrm{i} D^{(s)}_{\alpha\beta}\left(\mathrm{i}\frac{\partial}{m}\right) \Delta(x-y;m) , \qquad (14.88)$$

where $\Delta(x)$ is the famous Jordan–Pauli distribution. In (14.88) one has to take the commutator for integer spin and the anticommutator for half-integer spin. Otherwise the noncausal distribution Δ_1 would appear, and the field would be nonlocal.

We leave it as an exercise to introduce also quantum versions of the other field types, discussed in Sect. 5.1. For instance, one finds for the Bargmann–Wigner fields instead of (14.88) the following result (dropping indices and using the obvious generalization of Dirac's $\bar\psi$)

$$[\psi(x), \bar\psi(y)]_\pm = \bigotimes_j \Big[\mathrm{i}\gamma^\mu_{(j)}\partial_\mu + m\Big] \Delta(x-y;m) . \qquad (14.89)$$

What we have done in this section is, I believe, the kings way to the quantum theory of free fields for arbitrary spin.

14.A Appendix: Some Key Points of Mackey's Theory

Mackey's important theorem, formulated in Sect. 14.4, is based on his theory of imprimitivity systems. Let me first describe the connection between unitary representations of $G = A \rtimes H$ and systems of imprimitivity of H based on $\hat{A}$.

Let $g \mapsto W_g$ be a unitary representation of G in a Hilbert space $\mathcal{H}$ and let $U = W|A$, $V = W|H$ be its restrictions to A and H, respectively. According to the SNAG theorem we have the spectral decomposition

$$U_a = \int_{\hat{A}} x(a)\,\mathrm{d}P(x) , \qquad (14.90)$$

where P is a unique projection-valued measure on $\hat{A}$. From $ha = hah^{-1}$ we conclude that

$$V_h U_a V_h^{-1} = U_{ha} , \qquad (14.91)$$

implying that

$$V_h P(E) V_h^{-1} = P(hE) \,, \tag{14.92}$$

for every Borel set E of $\hat{A}$. By definition, the pair (V, P) is a *system of imprimitivity* for H based on $\hat{A}$. (V is a representation of H and P a projection-valued measure of $\hat{A}$, such that (14.92) is satisfied.)

Conversely, given such a system of imprimitivity (V, P), (14.90) defines a unitary representation U of A. Setting

$$W_{ah} = U_a V_h \,,$$

we obtain, as a result of (14.91) (implied by (14.92)), a representation of G, leading to the original system of imprimitivity. One can show that W is irreducible if and only if the corresponding system of imprimitivity is irreducible (in an obvious sense). An analogous statement holds for the notion of equivalence [14.10, Lemma 9.23].

The main part of Mackey's theory is concerned with the classification and description of irreducible systems of imprimitivity. A major tool in achieving this is Mackey's description of cohomology classes of cocycles [14.10, Theorem 8.27]). This leads to a $1:1$ correspondence between such cohomology classes and equivalence classes of systems of imprimitivity. (The main results are stated in [14.10, Theorems 9.7, 9.11].) For transitive systems of imprimitivity one then obtains a description in terms of representations of the stability group [14.10, Theorems 9.12, 9.20]. These results imply, in particular, Mackey's important theorem cited in Sect. 14.4.

The theory has, however, other interesting applications. It provides, for instance, a transparent uniqueness proof for the Schroedinger representation of the canonical commutation relations.

References

14.1 T. Damour: What is missing from Minkowski's "Raum und Zeit" lecture, Ann. Phys. **17**, 619–630 (2008), arXiv:0807.1300]

14.2 E. Wigner: Unitary representations of the inhomogeneous Lorentz group, Ann. Math. **40**, 149–204 (1939)

14.3 G.W. Mackey: Induced representations of locally compact groups, I, Ann. Math. **55**, 101–139 (1952)

14.4 G.W. Mackey: Induced representations of locally compact groups, II, Ann. Math. **58**, 193–221 (1953)

14.5 G.W. Mackey: Infinite dimensional group representations, Bull. Am. Math. Soc. **69**, 628–686 (1963)

14.6 N. Straumann: *Quantenmechanik: Ein Grundkurs über nichtrelativistische Quantentheorle* (Springer, Berlin 2002)

14.7 N. Straumann: *Relativistische Quantentheorie: Eine Einführung in die Quantenfeldtheorie* (Springer, Berlin 2005)

14.8 T. Frankel: *The Geometry of Physics: An introduction* (Cambridge Univ. Press, Cambridge 1997)

14.9 D.J. Simms: *Lie Groups and Quantum Mechanics*, Lecture Notes in Mathematics, Vol. 52 (Springer, Berlin 1968)

14.10 V.S. Varadarajan: *Quantum Theory of Covariant Systems*, Geometry of Quantum Theory, Vol. 2 (Van Nostrand, New York 1970)

14.11 U.H. Niederer, L. O'Raifeartaigh: Realizations of the unitary representations of the inhomogeneous Lorentz groups, Fortschr. Phys. **22**, 111–157 (1974)

14.12 S. Weinberg: Feynman rules for any spin, Phys. Rev. B **133**, 1318–1332 (1964)

14.13 S. Weinberg: Feynman rules for any spin II, Phys. Rev. B **134**, 882–896 (1964)

14.14 M. Fierz, W. Pauli: On the relativistic wave equations for particles of arbitrary spin in an electromagnetic field, Proc. R. Soc. A **173**, 211–232 (1939)

14.15 J. Mehra (Ed.): *The Physicist's Conception of Nature* (Reidel, Dortrecht, Boston 1973) p. 264

14.16 M. Reed, B. Simon: *Methods in Modern Mathematical Physics*, Vol. II (Academic, New York 1981), Chap. X

Part C

Spacetim

Part C Spacetime Structure and Mathematics

15. Spinors

Robert Geroch

Starting from an abstract complex 2-dimensional vector space with a fixed alternating tensor, there is constructed what is called a spinor space. This spinor space, it turns out, is intimately connected to what is known as a Lorentz vector space – a 4-dimensional vector space endowed with a metric of Lorentz signature. Finally, this connection between these two kinds of spaces is exploited to introduce, on virtually any spacetime, spinor fields.

Spinor fields, in many ways, merely reflect tensor fields. Every tensor field can be expressed in terms of one or more spinor fields; and the derivative operator on tensor fields extends uniquely to one on spinor fields. Thus, every algebraic calculation and every differential equation involving tensor fields has a direct spinor analog. It turns out, however, that for a number of topics – though by no means for all – the spinor version is simpler and more transparent than the tensor version. Examples include the classification of the Weyl tensor, the structure of the Maxwell field and the properties of null geodetic congruences. In addition, there are some topics for which spinors seem to be essential. These include the spin-s fields (for s a half-integer) and the Witten proof of positivity of gravitational mass. Acting on the spinor space is a certain group, the spinor group. The spinors generate the representations of that group, and in addition show how this group and its representations are related to those of the Lorentz group.

There are two common variants of spinors: 4-component spinors (which are used extensively in particle physics); and Euclidean spinors (which are used in, among other things, the Witten proof). There is also a number of subtleties involved in using spinors. For example: Some spacetimes admit no spinors at all, and for those that do admit a spinor structure it is not in general unique; there is in general not available any notion of the Lie derivative of a spinor field; and variational calculations involving spinors must be done with some care. Finally, there exist entire computational schemes – the spin-coefficient formalisms – based on spinors.

Spinors play two distinct roles in relativity.

On the one hand, there is a number of topics for which spinors seem to be essential. Examples include the description of Fermions in spacetime (e.g., via the Dirac equation); and the Witten proof that certain spacetimes have positive total mass, measured at spatial infinity. We simply do not have a viable description of Fermions, nor a version of the Witten proof, that circumvents spinors. Note that these topics are about physics, not just mathematics. The mere existence of such topics suggests that there is something fundamental about spinors. Also, since the complex numbers are

an integral part of spinors, this further suggests that the complexes are also somehow intertwined with the physical structure of spacetime.

The other role involves the use of spinors to carry out calculations and to illuminate the structure of various geometrical objects. Examples include the classification of the the Weyl tensor; and the derivation of the properties of shear-free null geodetic congruences. Spinors are not necessary for these tasks: They can be done perfectly well directly with tensors. But when spinors are useful, they are often *very* useful. For example, it is, in my opinion, easier first to learn spinors and then to learn the associated spinor classification of the Weyl tensor, than it is to learn the pure-tensor version of that classification. On the other hand, there are many other topics, such as the equations that describe a perfect fluid, for which spinors provide no advantage at all.

In both of these roles, spinors have the character of a "service subject" – rather like group theory or differential geometry. That is, spinors are important primarily because of the light they shed on other topics. To "learn" spinors is mostly to learn technique: How to calculate with them quickly and efficiently. This is not something to be absorbed passively, by merely reading. You must push some indices around for yourself.

The purpose of this chapter is to explain what spinors are, how they work, and how they are used. Sections 15.1 and 15.2 contain basic background material. We define spinor fields on a spacetime, discuss the relation between spinors and tensors, and give a few examples of how to do calculations with spinors. The remaining seven sections are a mixed bag of various topics, essentially independent of each other. These provide some examples of how spinors work (spinor structure, Lie derivatives, 4-spinors) and what they can be used for (representations of groups, Euclidean spinors, spin coefficients, variational problems). For other discussions of spinors, see [15.1–4].

Throughout this chapter, all manifolds are taken to be connected. Also, smoothness is assumed for every object for which "smooth" makes sense, e.g., for manifolds, maps, fields, etc.

15.1 Spinor Basics

In this section, we introduce spinor fields on spacetimes; as well as the basic operations available on such fields, the relation between spinor and tensor fields, etc.

Fix a complex, two-dimensional, vector space V.

We associate, with this V, three additional vector spaces, as follows. The first is the *dual* of V, written V^*. This, the complex vector space of all linear maps from V to the complexes, indeed has the structure of a vector space: We add such linear maps, and multiply them by complex numbers, in the obvious way. The second is the *complex conjugate* of V, written $\overline{V}$. As a set, $\overline{V}$ is the same as V. We keep this straight by means of the following notation. For $\alpha \in V$, we write $\overline{\alpha}$ for the corresponding element of $\overline{V}$. Addition and scalar multiplication in this set $\overline{V}$ are now defined, in terms of the corresponding operations in V, by the formula $c\overline{\alpha} + \overline{\beta} = (\overline{\overline{c}\alpha + \beta})$, where α, β are any two elements of V, and c is any complex number. In this formula, addition and scalar multiplication on the right-hand side are in V (where these operations are already defined); on the left, in $\overline{V}$ (where these operations are being defined). Here, $\overline{c}$ denotes the complex conjugate of this complex number. In other words, to take a linear combination of elements of $\overline{V}$, take the same linear combination, in V, of the corresponding elements of V, but apply complex conjugation to the scalars, as indicated above. It is easy to check that this set $\overline{V}$, with these operations, is indeed a vector space. The third vector space constructed from V is the complex conjugate of the dual of V (or, what is the same thing, the dual of the complex conjugate), written $\overline{V^*}$ (or $\overline{V}^*$).

Thus, we end up with a total of four complex, two-dimensional vector spaces, V, V^*, $\overline{V}$, and $\overline{V^*}$. Note that taking additional duals, or additional complex conjugates, on the vector spaces in this list simply returns other vector spaces in the list. The point of this construction is, not the vector spaces themselves (for if you have seen one complex, two-dimensional vector space, you have seen them all), but rather how these vector spaces are connected with each other. The connections between V, V^*, $\overline{V}$, and $\overline{V^*}$ are expressed in terms of two operations. First, each of V and V^* acts linearly on the other – and each of $\overline{V}$ and $\overline{V^*}$ also acts linearly on the other. This operation reflects the construction of the dual: Vector spaces dually related to each other act on each other. Second, there is given an antilinear isomorphism between V and $\overline{V}$ – and also one between V^* and $\overline{V^*}$. This operation reflects the construction of the complex conjugate: The complex-conjugation operation is an antilinear map.

We have introduced these four vector spaces, and the operations between them, by "building them up", starting from a single vector space V. But once the vector spaces have been constructed and the operations introduced, it is better to think of all four as being on the same footing.

It is convenient to introduce standard index notation. Elements of V will be denoted by upper-case Latin superscripts (e.g., ι^A); elements of V^* by subscripts (e.g., μ_C); elements of $\overline{V}$ by primed superscripts (e.g., $\kappa^{S'}$); and elements of $\overline{V^*}$ by primed subscripts (e.g., $\tau_{D'}$). That is, the indices tell us in which vector space an element resides. The operations relating these vector spaces are now expressed in terms of indices as follows. The action of elements of the dual of a vector space on a vector space (i. e., of V^* on V; of V on V^*; of $\overline{V}^*$ on $\overline{V}$; and of $\overline{V}$ on $\overline{V}^*$) is expressed by simply writing the two elements, with the same index, next to each other. Thus, $\tau = \mu_{C'}\rho^{C'}$ means: apply the element μ of $\overline{V}^*$ (the dual of $\overline{V}$) to the element ρ of $\overline{V}$, and denote the resulting complex number by τ. The operations of complex conjugation (from V to $\overline{V}$; from $\overline{V}$ to V; from V^* to $\overline{V^*}$; and from $\overline{V^*}$ to V^*) are all denoted by an overline. Thus, for $\beta_{K'}$ an element of $\overline{V^*}$, we write $\alpha_K = \overline{\beta_{K'}}$ or $\alpha_K = \overline{\beta}_K$ to mean: Apply the complex-conjugation operation to this β, and denote the resulting element of V^* by α. As an example of the notation, we have the following fact: $\mu_{A'}\overline{(c\nu^A)} = \overline{c}(\overline{\overline{\mu}_A \nu^A})$.

Next, we introduce all possible tensor products between these four vector spaces, possibly with repetitions. These are represented in standard index notation. That is, elements of tensor-product spaces are represented by the appropriate combinations of primed or unprimed subscripts or superscripts, where we agree that no repeated indices are to appear in any such tensor-product elements. Thus, for example, $\tau^{A'}{}_C$ would denote an element of $\overline{V} \otimes V^*$. Furthermore, the tensor product of two elements (having no index letter in common) is represented by writing the elements in juxtaposition. Thus, for example, the formula $\tau^{A'}{}_C = \alpha^{A'}\kappa_C$ means that τ is the tensor product of the elements $\alpha \in \overline{V}$ and $\kappa \in V^*$.

On these tensor-product elements, we have, by virtue of their construction, four operations: i) addition (applicable to two elements having precisely the same index structure); ii) outer product (take the tensor product of two elements, written, e.g., as $\rho^B{}_{D'E}\eta_S{}^{T'}$); iii) contraction (written using a repeated index, e.g., from $\rho^A{}_{D'B}$ in $V \otimes \overline{V}^* \otimes V^*$, we obtain $\rho^B{}_{D'B}$, an element of $\overline{V}^*$); and iv) complex conjugation (e.g., from $\tau^{A'}{}_C$ in $\overline{V} \otimes V^*$, we obtain $\overline{\tau}^A{}_{C'}$, an element of $V \otimes \overline{V}^*$). We regard complex numbers as elements of the tensor product of "no V's"; and thus scalar multiplication as a special case of ii). These four operations satisfy a very long list of properties: addition and outer product are associative; addition is commutative and associates over outer product; complex-conjugation, applied twice in succession, returns the original element; contraction commutes with addition, outer product and complex conjugation; outer product commutes with complex conjugation; etc.

Note that these tensor-product elements have many "interpretations". Thus, element $\tau^{A'}{}_C$ could be thought of, not only as an element of $\overline{V} \otimes V^*$ as above, but also as a linear map from V to $\overline{V}$ (with action $\alpha \to \tau^{A'}{}_C\alpha^C$); as a linear map from $\overline{V}^*$ to V^* (with action $\beta \to \tau^{A'}{}_C\beta_{A'}$, as an element of $(\overline{V}^* \otimes V)^*$ (with action $\gamma \to \tau^{A'}{}_C\gamma_{A'}{}^C$); as an element of $\overline{V \otimes \overline{V}^*}$; etc. We denote by $\delta^A{}_B$ the identity, i. e., the element satisfying $\delta^A{}_B\alpha^B = \alpha^A$ for every α^A. Also, we use round brackets to denote symmetrization (e.g., $\alpha_{(A'B')} = (1/2)(\alpha_{A'B'} + \alpha_{B'A'})$); and square brackets to denote antisymmetrization (e.g., $\alpha_{[A'B']} = (1/2)(\alpha_{A'B'} - \alpha_{B'A'})$).

The above is basic linear algebra – applicable quite generally to any finite-dimensional, complex vector space V. We now specialize to the structure of interest. Fix a nonzero, antisymmetric element, $\epsilon_{AB} = \epsilon_{[AB]}$, over V. Since V is two-dimensional, such an element always exists, and any two are related by a nonzero complex factor. We next introduce its inverse, ϵ^{AB}, satisfying $\epsilon^{AM}\epsilon_{BM} = \delta^A{}_B$, its complex conjugate, $\overline{\epsilon}_{A'B'} = \overline{\epsilon_{AB}}$, and its inverse-complex conjugate, $\overline{\epsilon}^{A'B'}$. These ϵ's will be used to raise and lower the indices, in the following manner. Given α^A and β_B, we set $\alpha_A = \alpha^M\epsilon_{MA}$ and $\beta^B = \epsilon^{BM}\beta_M$. Similarly for objects with primed indices, and with many indices. Note that raising and then lowering an index of a spinor returns that original spinor. While this convention for raising and lowering indices is extremely useful, one must take care to get the signs right. For example, we have $\alpha^A\beta_A = -\alpha_A\beta^A$. (Perhaps there could be invented some better notation.) Note that, for any $\alpha_{A'B'}$, we have $\alpha_{[A'B']} = (1/2)\alpha_{M'}{}^{M'}\overline{\epsilon}_{A'B'}$.

By a *spinor space*, we mean a complex two-dimensional vector space V, together with a nonzero antisymmetric element ϵ_{AB} of $V^* \otimes V^*$. Elements of the various tensor-product spaces constructed from a spinor space are called *spinors*. Thus, the (algebraic) operations on spinors are addition, outer product, contraction,

complex conjugation, and the raising and lowering of indices.

The key reason for interest in spinor spaces stems from the following construction. Fix a spinor space, (V, ϵ_{AB}). Denote by K the collection of all spinors, $\xi^{AA'}$, that are self-adjoint: $\overline{\xi^{AA'}} = \xi^{AA'}$. Then any real linear combination of elements of K is again in K, i. e., K acquires the structure of a real vector space. We next introduce an inner product on this K as follows: $\langle \eta, \xi \rangle = \eta^{AA'} \xi_{AA'}$. (Note that indices were lowered on the right-hand side.) This is clearly a real, symmetric inner product. One checks (e.g., by choosing a basis for V – more on this later) that this vector space K has dimension 4; and that the inner product $\langle\ ,\ \rangle$ has signature $(+, -, -, -)$. In short, we have constructed, from certain spinors and certain spinor operations, a *Lorentz vector space*.

Now fix any Lorentz vector space (T, g_{ab}). That is, T is a real, four-dimensional vector space, and g_{ab} is a metric over T of signature $(+, -, -, -)$ (which we may interpret as an inner product on T). We use standard (lower case Latin) indices for tensors over T, and we use the metric g_{ab} of T to raise and lower those indices. By a *spinor structure* on (T, g_{ab}), we mean a spinor space (V, ϵ_{AB}), together with an isomorphism between the vector space K (constructed above) and T that sends the inner product $\langle\ ,\ \rangle$ on K to that of the metric g_{ab} on T. Clearly, every Lorentz vector space admits a spinor structure. Furthermore, all spinor structures are obtained from any one by applying a suitable Lorentz transformation. It is convenient to describe a spinor structure by expressing the mapping in terms of a tensor $\sigma^b{}_{AA'}$ (i. e., in $T \otimes V^* \otimes \overline{V}^*$): The action of this mapping on $\xi^{AA'} \in K$ produces $\sigma^b{}_{AA'} \xi^{AA'} \in T$. The metric-preserving property can now be written in either (equivalent) form

$$\sigma^b{}_{AA'} \sigma^c{}_{DD'} g_{bc} = \epsilon_{AD} \overline{\epsilon}_{A'D'} \,, \tag{15.1}$$

$$\sigma^b{}_{AA'} \sigma^c{}_{DD'} \epsilon^{AD} \overline{\epsilon}^{A'D'} = g^{bc} \,. \tag{15.2}$$

Thus, the inverse map, from T back to K, is represented by $\sigma_b{}^{AA'}$.

Much of the usefulness of spinors arises from the interplay between K and T. This will be discussed extensively in the following section; but we give just one example here. Let o^A be any spinor. Then $o^A \overline{o}^{A'}$ is in K, and so $l^b = \sigma^b{}_{AA'} o^A \overline{o}^{A'}$, its image under the isometry, is in T. From (15.1), we have $l^a l_a = (o^A \overline{o}^{A'})(o_A \overline{o}_{A'})$. But by antisymmetry of ϵ_{AB}, there follows that $o^A o_A (= o^A o^B \epsilon_{BA}) = 0$, and so that the right-hand side of this expression vanishes. That is l^a is a null vector in T. It is easy to check that every, suitably directed, null vector can be written in this form; and that l^a determines o^A uniquely up to phase.

The object $\sigma^b{}_{AA'}$ above is called a *soldering form*, for it "solders" K to T. As we noted above, for fixed (V, ϵ_{AB}) and (T, g_{ab}) there are many such soldering forms. In some situations (e.g., for certain variational problems), one is contemplating dealing with a variety of such forms, and in those cases it is appropriate to retain $\sigma^b{}_{AA'}$, explicitly, in every equation. A much more common situation, however, is that in which one has a single soldering form $\sigma^b{}_{AA'}$, fixed throughout the discussion. When this is the case, it is convenient to suppress σ entirely. This is accomplished by the following notation: Instead of $\xi^b = \sigma^b_{AA'} \tau^{AA'}$, we write simply $\xi^b = \tau^{BB'}$, i. e., we think of "b" as merely standing for "BB'" (the σ that generates the transition between these two being understood). Also, similarly, we write $\kappa_c = \rho_{CC'}$ instead of $\kappa_c = g_{cb} \sigma^b{}_{AA'} \rho_{DD'} \epsilon^{AD} \overline{\epsilon}^{A'D'}$. In this notation we have, for example, $g_{ab} = \epsilon_{AB} \overline{\epsilon}_{A'B'}$.

This completes our discussion of the algebra of spinors. We have the notion of a spinor space, of spinors and the operations on them, of spinor structure on a Lorentz vector space, and of the soldering form, which represents that structure. The next step is to install spinors on spacetimes.

Fix a spacetime, (M, g_{ab}), so M is a 4-manifold, and g_{ab} is a Lorentz-metric field on this manifold. At each point p of M, the tangent space at p has the structure of a Lorentz vector space. By a *spinor structure* on (M, g_{ab}), we mean a (smooth) assignment of a spinor structure to each tangent space. A spinor structure on a spacetime, in other words, consists of two things. First, we must attach, to each point of M, a complex, two-dimensional vector space with alternating tensor. Second, we must specify a soldering-form field, $\sigma^a{}_{BB'}$, which, at each point of M, maps the self-adjoint spinors at that point to tangent vectors at that point. The question of which spacetimes admit any spinor structure at all, and, when there is such a structure, of how unique that structure is, will be discussed in a later section.

Fix a spacetime with spinor structure. Then, associated with each point of M, we have a spinor space, and so we can introduce the spinors of various ranks at that point. This can be done at each point of M, and so we have the notion of (smooth) spinor fields on M. One such field, for example, is ϵ_{AB}. The various operations on spinors at a point then extend to

corresponding operations on these spinor fields. Thus, we can add spinor fields (when they have precisely the same index structure), take outer products of spinor fields, and apply contraction and complex conjugation to spinor fields. Furthermore, we may raise and lower spinor indices (using ϵ_{AB} and its progeny), and translate between tensor and spinor fields (using the soldering form $\sigma^b{}_{AA'}$.) By taking further tensor products, we may also introduce mixed fields – having both tangent-space and spinor indices. One such field, for example, is $\sigma^b{}_{AA'}$ itself.

The above completes our discussion of the algebraic structure of spinor fields on spacetimes. We now turn to the differential structure. Fix a spacetime, (M, g_{ab}). Then, as is well known, there is one and only one (derivative) operator, ∇_a, acting on tensor fields on that spacetime, that i) is additive, commutes with contraction, and satisfies the Leibnitz rule under outer product; and ii) annihilates the metric: $\nabla_a g_{bc} = 0$. We shall always assume that such operators produce the gradient when applied to scalar fields, and satisfy $\nabla_{[a}\nabla_{b]}\phi = 0$ for any scalar field ϕ (i. e., are torsion-free).

It would also be extremely useful to be able take derivatives, not only of tensor fields, but also of spinor fields. That is, it would be useful to extend the action of this operator ∇_a to include also spinor fields. It turns out that there always exists a unique such extension.

Theorem 15.1
Fix a spacetime with spinor structure. Then there exists one and only one extension of ∇_a to spinor fields on that spacetime, that i) is additive, commutes with contraction and complex conjugation, and satisfies the Leibnitz rule under outer products; and ii) annihilates ϵ_{AB} and $\sigma^a{}_{BB'}$.

That is, the ∇_a of the theorem is to be applicable to both tensor fields and spinor fields (and, therefore, to mixed fields). Then the requirement (in ii) that $\nabla_c\sigma^a{}_{BB'} = 0$ guarantees that this derivative commutes with converting between tensors and spinors.

To prove this, fix any spinor field α_A in this spacetime, and set $F_{ab} = \alpha_A\alpha_B\bar{\epsilon}_{A'B'} + \epsilon_{AB}\bar{\alpha}_{A'}\bar{\alpha}_{B'}$. Then the real, antisymmetric tensor field F_{ab} is null, i. e., satisfies $F^{mn}F_{mn} = 0$ and $F_{[ab}F_{cd]} = 0$. It follows that $\nabla_k F_{ab}$ (where here ∇_k is the derivative operator on tensor fields) satisfies $(\nabla_k F_{mn})F^{mn} = 0$ and $(\nabla_k F_{[ab})F_{cd]} = 0$. But this, in turn, implies that $\nabla_k F_{ab} = 2\mu_{k(A}\alpha_{B)}\bar{\epsilon}_{A'B'} + 2\epsilon_{AB}\bar{\mu}_{k(A'}\bar{\alpha}_{B')}$ for some (clearly unique) field μ_{kA}. We now define an operator, ∇_k, on one-index spinor fields α_A by: $\nabla_k\alpha_A = \mu_{kA}$. Next, extend the action of this ∇_a to other one-index spinor fields using commutativity with contraction and complex conjugation, and then to arbitrary spinor fields using the Leibnitz rule. Finally, check that the resulting operator satisfies all the properties of the theorem. Uniqueness is clear from this construction.

For clarity, we made the hypothesis of this theorem stronger than necessary. One weakening that might seem attractive would be to remove the condition $\nabla_c\epsilon_{AB} = 0$. But this does not work: The uniqueness of ∇_c then fails. The derivative operator can also be written as $\nabla_{CC'}$ (i. e., $\sigma^b{}_{CC'}\nabla_b$, before suppression of the soldering form σ). Then the torsion-free condition, for example, becomes simply $\nabla_{(A}{}^{M'}\nabla_{B)M'}\phi = 0$.

Thus, on a spacetime with spinor structure we have spinor fields of various types: the ability to pass from tensor to spinor fields; the standard algebraic operations on spinor fields; and derivatives of those spinor fields.

It should be noted that at no point in the treatment above did we introduce frames of any kind. This is an important point, which is sometimes overlooked. The introduction, manipulation, and application of spinors *nowhere* requires any choice of frames, tetrads or bases.

15.2 Manipulating Spinors

To use spinors effectively, one must develop some facility for manipulating them and for passing back and forth between tensors and spinors. In this section, we provide a few examples of such calculations. While these examples are far from illustrating everything that can be done with spinors, they will, I hope, at least give a sense of what is involved.

Fix, once and for all, a spacetime (M, g_{ab}) and a spinor structure on that spacetime. Let us consider, in this spacetime, a Maxwell field, F_{ab}. That is, F_{ab} is antisymmetric, and satisfies Maxwell's equations.

The first step is to turn the tensor field F_{ab} into a spinor field using the soldering form. There results $F_{AA'BB'} = \sigma^m{}_{AA'}\sigma^n{}_{BB'}F_{mn}$. Next note that any spinor involving two indices, A and B, is equal to the sum of that spinor symmetrized in these two indices and that spinor antisymmetrized. But the latter is a multiple of ϵ_{AB}. Applying this fact in the present case, we obtain

$F_{AA'BB'} = \alpha_{AA'BB'} + \beta_{A'B'}\epsilon_{AB}$, where $\alpha_{AA'BB'}$ is symmetric in "A, B". Repeating the same procedure on the primed indices of α and β, we obtain

$$F_{AA'BB'} = \kappa_{AA'BB'} + \tau_{A'B'}\epsilon_{AB} + \phi_{AB}\overline{\epsilon}_{A'B'} + \rho\epsilon_{AB}\overline{\epsilon}_{A'B'} \,, \tag{15.3}$$

where τ and ϕ are each symmetric spinors, and κ is symmetric in both index pairs "A, B" and "A', B'". But F_{ab} is an antisymmetric tensor, and so $F_{AA'BB'}$ must reverse sign under simultaneous interchange of index pairs "A, A'" and "B, B'". This implies that κ and ρ above both vanish. Finally, reality of the Maxwell tensor F_{ab} implies that $\tau = \overline{\phi}$. We conclude that our Maxwell field F_{ab} can be written uniquely as

$$F_{AA'BB'} = \phi_{AB}\overline{\epsilon}_{A'B'} + \overline{\phi}_{A'B'}\epsilon_{AB} \,, \tag{15.4}$$

where ϕ_{AB} is symmetric in its two indices. Thus, the skew tensor F_{ab} becomes, in spinor form, a symmetric spinor ϕ_{AB}. This ϕ_{AB} is called the *Maxwell spinor.* Note that the dimensions work out: The vector space of skew tensors F_{ab} at a point has dimension 6, the same as the (real) dimension of the space of symmetric, second-rank spinors. The Maxwell spinor, then, simply reorganizes the components of the Maxwell tensor in a different way.

The idea, now, is to re-express various algebraic features of Maxwell fields in terms of the Maxwell spinor. First note that the alternating tensor of (M, g_{ab}) is given by

$$\epsilon_{abcd} = \mathrm{i}(\epsilon_{AB}\epsilon_{CD}\overline{\epsilon}_{A'C'}\overline{\epsilon}_{B'D'} - \overline{\epsilon}_{A'B'}\overline{\epsilon}_{C'D'}\epsilon_{AC}\epsilon_{BD}) \,. \tag{15.5}$$

This follows, noting that the right-hand side is indeed antisymmetric in "a, b, c, d", and satisfies $\epsilon^{abcd}\epsilon_{abcd} = -24$. For the former, it suffices to check: If we symmetrize over any two unprimed indices in (15.5), then the result is antisymmetric in the corresponding two primed indices. From (15.5), it follows that the spinor version of the dual of F, ${}^*F_{ab} = (1/2)\epsilon_{ab}{}^{mn}F_{mn}$ is simply $\mathrm{i}\phi_{AB}$. That is, dualization on skew tensors is multiplication by i on the corresponding spinors. By direct computation, we now find

$$\phi^{AB}\phi_{AB} = 1/4(F^{ab}F_{ab} + \mathrm{i}F^{ab*}F_{ab}) \,. \tag{15.6}$$

Thus, the two (real) scalar invariants of the Maxwell field are combined into a single complex scalar invariant constructed directly from the corresponding Maxwell spinor. For the stress–energy of the Maxwell field, we have

$$T_{ab} = F_a{}^m F_{bm} - 1/4 g_{ab}F^{mn}F_{mn} = \phi_{AB}\overline{\phi}_{A'B'} \,. \tag{15.7}$$

Several facts about the Maxwell stress-energy are immediate from this expression. First, the vanishing of the trace of the stress-energy, $0 = g^{ab}T_{ab} = \epsilon^{AB}\overline{\epsilon}^{A'B'}T_{AA'BB'}$, follows from symmetry of ϕ_{AB}. Second, the stress-energy of F_{ab} is the same as the stress-energy of its dual. This follows, replacing ϕ_{AB} by $\mathrm{i}\phi_{AB}$ in (15.7). The energy condition on the stress-energy arises as follows. First note that a vector t^a is timelike, with the same time-orientation as the null vectors $l^a = \alpha^A\overline{\alpha}^{A'}$, if and only if $t^a l_a >$, i. e., if and only if the quadratic form $t^{AA'}\alpha_A\overline{\alpha}_{A'}$ is positive-definite. The energy condition now follows directly, contracting (15.7) with $t^{AA'}t^{BB'}$ and using positive-definiteness. Finally, we note, from (15.7), that $T_a{}^m T_{bm}$ is a multiple of g_{ab}. This is far from obvious from the tensor expression for the stress-energy.

Recall that a Maxwell field is called *null* provided both of its scalar invariants vanish. In spinor terms, this means that $\phi^{AB}\phi_{AB} = 0$. But this equation implies that $\phi_A{}^M\phi_{BM} = 0$ (since the left-hand side of this is already antisymmetric in "A, B"); which in turn implies that ϕ_{AB}, regarded as a linear map from one-index spinors to one-index spinors, has rank one; which, finally, implies that $\phi_{AB} = \alpha_A\alpha_B$ for some one-index spinor α_A. The converse – that every Maxwell spinor is of this form is null – is immediate. Now set $l^a = \alpha^A\overline{\alpha}^{A'}$, a null vector. It follows, from (15.4), that $l^a\, F_{ab} = 0$. Thus, a null Maxwell field annihilates a nonzero null vector (and conversely). Note also, from (15.7) that, for a null Maxwell field, $T^{ab} = l^a l^b$. These are, of course, well-known facts about Maxwell fields. But sometimes things become more transparent when expressed using spinors.

It is an important fact – which we shall use later – that a null F_{ab} determines the spinor α_A uniquely up to sign.

More generally, every Maxwell spinor can be written in the form $\phi_{AB} = \alpha_{(A}\beta_{B)}$, for some spinors α_A and β_B, where these spinors are unique up to exchanging a complex factor. To see this, perform in turn the following four steps: choose a basis, o^A, ι^A, for spinor space; consider the quadratic polynomial in a complex variable z given by $\phi_{AB}(o^A + z\iota^A)(o^B + z\iota^B)$; factor this polynomial as a product of two factors linear in z; and write each of the two factor as some one-index spinor, α_A and β_B, contracted with $(o^A + z\iota^A)$. The spinors α, β so defined are called the *principal spinors* of the

Maxwell spinor. Define, from the principal spinors, two null vectors $l^a = \alpha^A \overline{\alpha}^{A'}$, $n^a = \beta^A \overline{\beta}^{A'}$. Substituting from (15.4), we find that $l_{[a}F_{b]m}l^m = 0$ and $n_{[a}F_{b]m}n^m = 0$. These will be recognized as the defining equations for the principal null directions of the Maxwell tensor. Thus, the principal spinors are simply the spinor renditions of the principal null vectors.

We now turn to Maxwell's equations, which we may write as $\nabla^m F_{am} = 0$ and $\nabla^m {}^*F_{am} = 0$. Substituting (15.4) into the former, we obtain

$$\nabla^A{}_{B'}\phi_{AB} + \nabla_B{}^{A'}\overline{\phi}_{A'B'} = 0\,. \tag{15.8}$$

The other Maxwell equation yields the same formula, but with the sign of the second term reversed. Thus, the entirety of Maxwell's equations becomes, in spinor form

$$\nabla^A{}_{B'}\phi_{AB} = 0\,. \tag{15.9}$$

If there were sources, then on the right-hand side there would appear a vector, $J_{BB'}$. The real and imaginary parts of this (in general complex) vector are the electric and magnetic charge-currents, respectively. Applying $\nabla^{AA'}$ to the stress-energy (15.7), and using Maxwell's equations, (15.9), without sources, we see immediately that the stress-energy is conserved.

Consider next Maxwell's equations in the null case. Substituting $\phi_{AB} = \alpha_A\alpha_B$ into (15.9), and contracting with $\alpha^B\overline{\alpha}^{B'}$, we obtain $\alpha^B\alpha^A\overline{\alpha}^{B'}\nabla_{AB'}\alpha_B = 0$. Setting $l^a = \alpha^A\overline{\alpha}^{A'}$, this can be rewritten as $\alpha^B l^a\nabla_a\alpha_B = 0$. But this means that $l^a\nabla_a\alpha_B$ is a multiple of α_B, which in turn implies that $l^a\nabla_a l_b$ is a multiple of l_b. We have proved: For a null solution of Maxwell's equations, the null vector field l^a is geodetic.

Many of the features we have seen in the Maxwell case above are merely examples of more general facts about spinors.

Every spinor can be written in terms of spinors that are totally symmetric in all unprimed indices, and also totally symmetric in all primed indices. To see this, consider a spinor $\alpha_{A\cdots D}$, with n indices. First note that every rearrangement of the indices of α can be obtained from the arrangement $\alpha_{A\cdots D}$ by switching indices two at a time. Therefore, every such rearrangement differs from $\alpha_{A\cdots D}$ by terms involving ϵ_{EF} and spinors of lower rank. It follows that $\alpha_{(A\cdots D)}$ differs from $\alpha_{A\cdots D}$ by such terms. In other words, $\alpha_{A\cdots D}$ is equal to $\alpha_{(A\cdots D)}$ plus terms involving lower-ranked spinors. Now repeat this procedure for the spinors of lower rank; and continue in this way until all spinors have been replaced by totally symmetric spinors. The result – that every spinor can be written in terms of spinors separately totally symmetric in primed and in unprimed indices – now follows, by the same argument.

In many cases (such as the Maxwell case), this decomposition of spinors into totally symmetric spinors reveals interesting features. But in other cases it reveals nothing at all. For instance, applied to the stress-energy tensor this decomposition yields $T_{ab} = T_{AA'BB'} = K_{AA'BB'} + 1/4T\epsilon_{AB}\overline{\epsilon}_{A'B'}$, where K is symmetric in "A, B" and "A', B'" and $T = T^m{}_m$.

Every totally symmetric spinor field, $\zeta_{A\cdots D} = \zeta_{(A\cdots D)}$, can be written as a symmetrized product of 1-index spinors: $\zeta_{A\cdots D} = \mu_{(A}\cdots\nu_{D)}$. This follows from the same argument as we gave for the Maxwell case. The spinors $\mu_A, \ldots, \nu_D$ are called the *principal spinors* of ζ. They are unique up to exchanging complex factors. In the Maxwell case, the Maxwell spinors, ϕ_{AB} fall into two classes: the null spinors, when the two principal spinors coincide, and the non-null, when they do not. One can carry out a similar classification for higher-rank (totally symmetric) spinors, resulting in more classes. For example, for a fourth-rank spinor ψ_{ABCD}, there are five classes: $\alpha_{(A}\beta_B\gamma_C\delta_{D)}$, $\alpha_{(A}\alpha_B\beta_C\gamma_{D)}$, $\alpha_{(A}\alpha_B\beta_C\beta_{D)}$, $\alpha_{(A}\alpha_B\alpha_C\beta_{D)}$, and $\alpha_A\alpha_B\alpha_C\alpha_D$.

A mass-zero *spin-s field* (where s is a positive half-integer) consists of a $2s$-index, totally symmetric spinor, $\phi_{AB\cdots D}$, satisfying the equation $\nabla^M{}_{A'}\phi_{MB\cdots D} = 0$. So, for example, the Maxwell field is a spin-1 field. Every spin-s field, then, gives rise to $2s$ principal spinors, some of which may coincide. In the case of a null Maxwell spinor (when the two principal spinors coincide) that spinor gives rise to a null, geodetic vector field. By the same argument, for a null spin-s field (all principal spinors coincide) with $s \geq 1$, there again arises a null, geodetic vector field. This is a portion of the Goldberg–Sachs theorem [15.5]. When s is an integer, we can reexpress the spin-s field in terms of a tensor, and rewrite the field equation as an equation on this tensor. But this quickly gets very complicated. Try it for $s = 3$. But when s is not a whole integer there is no simple tensor rendition of the spin-s field: Spinors seem to play an essential role in describing the physics of these particles.

In the case $s = 1/2$, we obtain the what is called the *Weyl neutrino* equation. On the other hand, a massive spin-1/2 particle is described by a pair of spinor fields,

(ξ^A, η'_A), satisfying the *Dirac equation*

$$\nabla_{AA'}\xi^A = \frac{m}{\sqrt{2}}\,\eta_{A'}\,, \qquad \nabla^{AA'}\eta_{A'} = -\frac{m}{\sqrt{2}}\xi^A\,, \quad (15.10)$$

where m is the mass. Taking a derivative of the first equation in (15.10), and using the second, we obtain $\nabla_b\nabla^b\xi^A = -m^2\xi^A$. That is, ξ^A (and, similarly, $\eta_{A'}$) satisfy, by virtue of the Dirac equation, the Klein–Gordon equation with mass m. It follows from the Dirac equation that the future-directed timelike-or-null vector field $\xi^A\overline{\xi}^{A'} + \eta^{A'}\overline{\eta}^A$ is conserved. This is the charge-current vector of the Dirac field.

Finally, we discuss briefly how curvature enters spinor calculations. Fix a spinor field α_C. It then follows, from the Leibnitz rule, that $\nabla_{[a}\nabla_{b]}\alpha_C$ is linear in α_C. Therefore, there exists a spinor field $\rho_{abC}{}^D$ such that

$$\nabla_{[a}\nabla_{b]}\alpha_C = \rho_{abC}{}^D\alpha_D\,. \quad (15.11)$$

This $\rho_{abC}{}^D$ is the "effective curvature" for spinor fields. Analogous formulae follow for spinors with a single primed index (using that ∇_a commutes with complex conjugation), for spinors with a single superscript (using that ∇_a commutes with contraction and satisfies the Leibnitz rule), and for spinors with many indices (again using commutation and Leibnitz). It follows, from the fact that $\nabla_a\epsilon_{BC} = 0$ that ρ_{abCD} is symmetric in indices "C, D". Now consider $\nabla_{[a}\nabla_{b]}\alpha_{CC'}$. We may evaluate this, in terms of ρ, using (15.11). Alternatively, since $\alpha_{CC'}$ is a merely a covariant vector field on the spacetime, we may also evaluate this expression in terms of the Riemann tensor. Equating these two, we obtain

$$R_{abcd} = \rho_{abCD}\overline{\epsilon}_{C'D'} + \overline{\rho}_{abC'D'}\epsilon_{CD}. \quad (15.12)$$

This equation says what one would have expected: Regard R_{abcd} as a "Maxwell field" in its last two indices, keeping the first two indices intact. Then the associated "Maxwell spinor" is just ρ. In any case, (15.12) shows that R_{abcd} and ρ_{abCD} carry precisely the same information.

The Riemann tensor, rewritten as a spinor, can be decomposed in terms of spinors that are totally symmetric, as noted earlier. The result is to split the Riemann tensor into three pieces: the scalar curvature R, the trace-free Ricci tensor, $R_{ab} - 1/4Rg_{ab}$, and the Weyl tensor, C_{abcd}. Spinors do not have much to say about the first two, but they do about the third. The spinor equivalent of the Weyl tensor is given by

$$C_{abcd} = \psi_{ABCD}\overline{\epsilon}_{A'B'}\overline{\epsilon}_{C'D'} + \overline{\psi}_{A'B'C'D'}\epsilon_{AB}\epsilon_{CD}\,. \quad (15.13)$$

Here, ψ_{ABCD}, the *Weyl spinor* is totally symmetric. Compare (15.4). Note that the dimensions work out: Total symmetric fourth-rank spinors have (real) dimension 10, the same as that of Weyl tensors. We remarked earlier about the general classification of totally symmetric spinors. Applied to the Weyl tensor, we recover [15.3] the five classes comprising the Petrov classification [15.6].

Let the Ricci tensor vanish, so the entire Riemann tensor is comprised of its Weyl part. Then the Bianchi identity becomes

$$\nabla_{A'}{}^A\psi_{ABCD} = 0\,. \quad (15.14)$$

That is, the Weyl tensor in this case is just a spin-2 field! Next, consider the tensor field $T_{abcd} = \psi_{ABCD}\overline{\psi}_{A'B'C'D'}$, constructed from the Weyl tensor. It is clear from its definition that this tensor field is totally symmetric, and positive definite, in the sense that $T_{abcd}t^at^bt^ct^d \geq 0$ for every timelike t^a. Furthermore, it follows from (15.14) that T_{abcd} is conserved: $\nabla^aT_{abcd} = 0$. This is the Bel–Robinson tensor [15.7].

15.3 Groups; Representations

Fix a Lorentz vector space, (T, g_{ab}), with spinor structure, $(V, \epsilon_{AB}, \sigma^b{}_{AA'})$.

The group of all ϵ-preserving isomorphisms on the vector space V is called the *spinor group*, S, of (V, ϵ_{AB}). In more detail, an element of S is a spinor $S^A{}_B$ (a linear map on V) satisfying $S^M{}_AS^N{}_B\epsilon_{MN} = \epsilon_{AB}$ (epsilon-preservation). The group operation in S is composition: $S\hat{S} \to S^A{}_M\hat{S}^M{}_B$. This group (also known as $SL(2, C)$) is six-dimensional: It is connected and simply connected, and its topology is the product of the 3-sphere, S^3, and R^3.

The group of all g_{ab}-preserving isomorphisms on the vector space T is called the *Lorentz group*, $\mathcal{L}$, of (T, g_{ab}). An element of $\mathcal{L}$ is a tensor, $L^a{}_b$, satisfying $L^m{}_aL^n{}_bg_{mn} = g_{ab}$. The group operation in $\mathcal{L}$ is also composition: $L\hat{L} \to L^a{}_m\hat{L}^m{}_b$. This group (also known

as $O(3,1)$) is also six-dimensional. It has four connected components (consisting, respectively, of Lorentz transformations that preserve both time- and space-orientation; those that preserve one and not the other; and those that preserve neither). Each component has topology the product of the projective 3-sphere RP^3 (i. e., the 3-sphere with opposite points identified) and R^3. Thus, each connected component of the Lorentz group is doubly connected. Those Lorentz transformations that preserve both orientations form a normal subgroup of $\mathcal{L}$ called the *restricted Lorentz group*. The quotient of the Lorentz group by the restricted Lorentz group is a group with four elements, isomorphic with $Z_2 \times Z_2$ (i. e., isomorphic with the group of all ways to "flip orientations"). The double connectivity of the restricted Lorentz group can be pictured as follows. Consider, in this group, the family of rotations, about some fixed spatial axis, through angles ranging from 0 to 2π. This is a closed curve in the restricted Lorentz group. It cannot be contracted to a point within that group. However, the closed curve that traverses this one twice (i. e., that consists of rotations through angles ranging from 0 to 4π), *can* be contracted to a point.

There is a natural Lie-group homomorphism from S to $\mathcal{L}$. It sends $S^A{}_B \in S$ to the $L^a{}_b \in \mathcal{L}$ given by $L^a{}_b = S^A{}_B \overline{S}^{A'}{}_{B'}$. Note that this is indeed a map from S to $\mathcal{L}$, and that it is indeed product- and inverse-preserving (i. e., that it is indeed a homomorphism of groups). This homomorphism is two-to-one: the two elements S and $-S$ of S are sent to the same element L of $\mathcal{L}$. This homomorphism is onto the restricted Lorentz group. In other words, every restricted Lorentz transformation $L^a{}_b$ can be written in the form $S^A{}_B \overline{S}^{A'}{}_{B'}$ with $S^A{}_B$ in S, and this $S^A{}_B$ is unique up to sign.

Every curve γ in the Lorentz group, starting at the identity, $\delta^a{}_b \in \mathcal{L}$, can be lifted uniquely to a curve γ' in S starting at the identity, $\delta^A{}_B$, of that group. Here "lifted" means that the image of γ' under the homomorphism $S \to \mathcal{L}$ is precisely γ. (This uniqueness follows from the fact that S is a covering manifold of the restricted Lorentz group.) Consider the "2π-rotation curve" described earlier. The lifting of this closed curve is a curve in S that joins the identity, $\delta^A{}_B$, to $-\delta^A{}_B$. That is, the lifting is not a closed curve. In follows that this closed curve cannot be contracted to a point in $\mathcal{L}$. However, the lifting of the 4π-rotation curve *is* closed in S; and this curve *can* be contracted to a point. All this is a reflection of the simple connectivity of S and the double connectivity of $\mathcal{L}$.

The remaining elements of the Lorentz group (i. e., those that reverse one or both of time- and space-orientation) are represented in terms of spinors as follows. Those that reverse both orientations are of the form $L^a{}_b = -S^A{}_B \overline{S}^{A'}{}_{B'}$, where again S is ϵ-preserving and again S is uniquely determined by L, up to sign. Those that preserve time-orientation but not space-orientation are of the form $L^a{}_b = K^A{}_{B'} \overline{K}^{A'}{}_B$, where $K^A{}_{B'}$ satisfies $K^M{}_{A'} K^N{}_{B'} \epsilon_{MN} = \overline{\epsilon}_{A'B'}$. Finally, those that preserve space orientation but not time orientation are given by the same formula, but with a minus sign on the right; for the same class of spinors $K^A{}_{B'}$. In these cases, again, L determines K uniquely up to sign. Thus, in particular, every element $L^a{}_b$ of the Lorentz group $\mathcal{L}$ can be expressed, spinorially, in one and only one of these four forms.

Fix an arrangement of spinor indices, say $()^{A'C}{}_{DF}$. Consider the complex vector space of all spinors having this index arrangement. Its (complex) dimension is 2 raised to the power of the number of indices (i. e., in this example, 16). The spinor group S acts on this vector space as follows: $S^A{}_B \in S$ sends $\alpha^{A'C}{}_{DF}$ to $\overline{S}^{A'}{}_{M'} S^C{}_N S_D{}^K S_F{}^L \alpha^{M'N}{}_{KL}$. This action is clearly a representation of the group S. In this way, then, we acquire a large number of representations of the spinor group.

The various spinor operations now become operations on these representations. Taking outer products of spinors corresponds to taking tensor products of representations. For example, the tensor product of the representation on spinors α^A and that on $\beta_{B'}$ is the representation on spinors $\tau^A{}_{B'}$. Taking complex conjugates corresponds to taking the complex-conjugate representation. Raising and lowering of indices provide isomorphisms between certain representations. Consider, for example, the representation on spinors α^A, and that on spinors β_A. "Raising and lowering the spinor index" provides an isomorphism between the underlying vector spaces, which (by virtue of the fact that each $S \in S$ leaves ϵ_{AB} invariant) commutes with the action of S. In other words, raising and lowering generates an isomorphism between these representations. A similar result holds, clearly, for spinors of other ranks. We may eliminate these equivalent representations from our list by restricting to those representations based on spinors with only subscripts (since any representation based on spinors having superscripts is equivalent to some subscript-only representation).

The fact that every spinor can be written in terms of symmetric spinors means that some of these rep-

resentations are direct sums of others. For example, every second-rank spinor α_{AB} can be written uniquely as $\beta_{AB} + \kappa\epsilon_{AB}$ with β_{AB} a symmetric spinor and κ a complex number; and, clearly, this decomposition is preserved by the action of elements of the spinor group. This means that the representation based on general spinors α_{AB} is the direct sum of the representation based on symmetric spinors β_{AB} and that based on scalars κ. A similar result holds for spinors with more indices. We conclude that every spinor representation of the spinor group S is a direct sum of representations on spinors having only subscripts, and totally symmetric in those subscripts. Consider the representation of the spinor group based on spinors of the form $\tau_{A\cdots CD'\cdots E'}$, where τ is totally symmetric in $A \cdots C$ and in $D' \cdots E'$. This is called the $(n/2, n'/2)$-representation, where n is the number of unprimed indices, and n' the number of primed indices, of τ. The (complex) dimension of the underlying vector space is $(n+1)(n'+1)$. We have shown, then, that every spinor representation of S is a direct sum of these $(n/2, n'/2)$-representations.

Take the tensor product of two of these symmetric-spinor representations, and then write the spinors that arise from the outer products back in terms of symmetric spinors. There results a formula, expressing the tensor product of two symmetric-spinor representations as a direct sum of symmetric-spinor representations. The coefficients in this expression are called the *Clebsch–Gordon coefficients.*

Thus, we have obtained, from these totally symmetric spinors $\tau_{A\cdots CD'\cdots E'}$, a list of representations of S. It turns out that these representations have just about every desirable property one could imagine. They are distinct (i. e., no two of these representations are isomorphic with each other) and irreducible (i. e., none can be written as any direct sum of any representations of lower dimension). More important, these exhaust the collection of all finite-dimensional representations of S, in the following sense: Every finite-dimensional representation of the spinor group S is isomorphic to some direct sum of these standard representations.

Much of the above applies also to representations of the Lorentz group. But this case is a little more complicated. The main difference is this: It is false that every tensor over a Lorentz vector space can be written in terms of symmetric tensors. But it is still possible to find a list of "fundamental tensor representations", where these are based on more complicated tensor symmetries. There are again formulae that express the tensor product of two of these as a direct sum of fundamental representations. These are again distinct and irreducible; and, again, every finite-dimensional representation of the Lorentz group is isomorphic to a direct sum of these.

15.4 Spinor Structure

Let (M, g_{ab}) be a spacetime. Recall that a *spinor structure* on (M, g_{ab}) consists of a smooth assignment of a spinor structure to the tangent space of each point of M. The procedure for constructing a spinor structure, in more detail, is the following. First, we must introduce a smooth fibre bundle over M, where the fibre at each point has the structure of a complex two-dimensional vector space with alternating tensor. Then, we must introduce on M a smooth soldering-form field, $\sigma^b{}_{AA'}$, which provides an isometry, at each point, between the tangent space of M at that point, and the self-adjoint spinors at that point. Which spacetimes have such spinor structures?

Call two spinor structures on spacetime (M, g_{ab}), *equivalent* if there exists, globally over all of M, an ϵ-preserving map between the spinor spaces that sends one soldering form to the other. For equivalent spinor structures, then, one can pass from one spinor space to the other in a way that preserves the relationship between spinors and tensors. In other words, equivalent spinor structures are functionally identical. Given a spacetime that has one spinor structure, how many, up to equivalence, does it have? And how can the various spinor structures be characterized?

These are the subjects of this section.

We first note that a spinor structure on (M, g_{ab}) automatically endows that spacetime with a specific time-orientation (namely, that for which the future light cones are those containing null vectors of the form $\alpha^A\overline{\alpha}^{A'}$); and also with a specific space-orientation (namely, that which arises from the time-orientation and the alternating tensor (15.5) generated from ϵ_{AB}). It follows from this remark that if a spacetime fails to be time-orientable or fails to be space-orientable then it cannot have any spinor structures at all. It further follows that a necessary condition that two spinor structures on a spacetime be equivalent is that they induce the same time-orientation and the same space-orientation on that spacetime.

Roughly speaking, in order that a spacetime (M, g_{ab}) have any spinor structure at all, it must, in addition to being time- and space-orientable, satisfy a certain topological condition. This condition is satisfied for the vast majority of spacetimes of physical interest. In fact, it is not even so easy to find an example of a spacetime on which it fails. Now suppose that a spacetime satisfies this condition, so it does have at least one spinor structure. Then in general there there will exist a number of inequivalent spinor structures. These are classified by making a specific choice of a time- and space-orientation for that spacetime, and, in addition making one further choice, involving the first homotopy group of the underlying manifold M. The details are as follows.

Let (M, g_{ab}) be a spacetime. Fix, once and for all, a separate Lorentz vector space, (T', g'). Consider any point p of M, and consider the tangent space, T_p at that point. This T_p, by virtue of the spacetime metric g_{ab}, itself has the structure of a Lorentz vector space. Denote by $\mathcal{F}_p$ the collection of all isometries between the two Lorentz vector spaces $(T_p, g_{ab}|_p)$ and (T', g'). Given one such isometry, then all of the other isometries are obtained from this one by composing it with the Lorentz transformations on (T', g'). Thus, the collection $\mathcal{F}_p$ of all these isometries is a copy of the Lorentz group, i. e., it is a 6-manifold, consisting of four connected pieces, each of which is diffeomorphic with the projective 3-sphere cross R^3. Next, consider the fiber bundle, B, over M whose fiber, at each point $p \in M$ is this $\mathcal{F}_p$. Thus, the bundle manifold of B is 10-dimensional (four dimensions being required to locate the point p of M; and then six more dimensions to locate the point of $\mathcal{F}_p$).

A *cross-section* of the fibre bundle B is a map that assigns (smoothly), to each point p of M, an element of $\mathcal{F}_p$. In other words, a cross-section assigns, to each point p of M, an isometry between the tangent space T_p at that point and our fixed Lorentz vector space (T', g'). Note that a cross-section is a global object: It makes these assignments over all of M. Now, a given spacetime (M, g_{ab}) may or may not admit any such cross-sections. If the spacetime is Minkowski, for instance, then clearly there are cross-sections. By contrast, if the given spacetime fails to be time-orientable, or fails to be space-orientable, then there are no such cross sections. This is easy to see: fix a time- and space-orientation on the Lorentz vector space (T', g'). Then a cross-section would induce, at each point p of M, a time- and space-orientation on the tangent space T_p at that point, i. e., would time- and space-orient (M, g_{ab}).

Next, fix, once and for all, a spinor structure on this Lorentz vector space (T', g'). That is, fix a complex, two-dimensional vector space V with alternating tensor ϵ, together with a soldering form σ to (T', g').

Now suppose for a moment that we have found some cross-section of the bundle B over M. Then we may construct from that cross-section a spinor structure on (M, g_{ab}). This construction is the obvious one: we take, as our spinor bundle over M, the simple product $M \times V$. Then this cross-section (which connects T' to each tangent space T_p), together with the soldering form on our spinor space (which connects V to T') yields a soldering form for this bundle, i. e., a soldering form field on M.

Thus, we have described one, very direct, way to obtain a spinor structure on a spacetime: Introduce the bundle B, find a cross-section of that bundle, and from that cross-section build the spinor structure as described above. Take M to be noncompact. (The compact case is a little different. But this case is also not very important, as virtually every physically interesting spacetime is noncompact.) Then this is not only *one* way to obtain a spinor structure – it is the *only* way. It is known [15.8] that *every* spinor structure on noncompact (M, g_{ab}) arises from a cross-section of the bundle B. But, as remarked earlier, not every time- and space-orientable spacetime admits [15.8] such a cross-section.

We next consider uniqueness. Suppose that we have two spinor structures constructed as above. That is, we have two cross-sections of the bundle B, each giving rise to its own spinor structure. Suppose further that these two cross-sections are homotopic to each other – i. e., that one cross-section can be continuously deformed, through a family of cross-sections, to the other. It then follows, we claim, that the two spinor structures are equivalent. To see this, fix a point p of M. Then each cross-section, evaluated at p, produces an element of $\mathcal{F}_p$. The homotopic family then gives rise to a curve in $\mathcal{F}_p$ with initial point the element, F_0, of $\mathcal{F}_p$ arising from one of the cross-sections; and final point the element F_1 arising from the other. This curve can be described by applying, to the element F_0, a curve in the Lorentz group of (T', g') with initial point the origin. But there is a unique lifting of this curve to a curve, again with initial point the origin, in the spinor group of (V, ϵ). The final point of this lifted curve gives a certain element S_0, of the spinor group of (V, ϵ). This S_0 provides the desired mapping between the two spinor spaces at p. Repeating this construction for each point p of M, we obtain an equivalence between the two spinor

structures. Note that the homotopic family of cross sections plays a crucial role in this argument. Without the homotopic family, we would still have a Lorentz transformation on (T', g') sending F_0 to F_1, but there are two elements of the spinor group associated with this element of the Lorentz group, and we would have no guarantee that there is a consistent way of resolving this sign ambiguity over all of M.

Thus, homotopic cross-sections give rise to equivalent spinor structures. Note that this is consistent with our earlier observations, for homotopic cross-sections certainly give rise to the same time- and space-orientations on (M, g_{ab}). Here, then, is one way to show equivalence of two spinor structures. It turns out (at least, when M is noncompact) that this is not only *one* way – it is the *only* way. It is known that two spinor structures are equivalent if and only if their associated cross-sections are homotopic.

We summarize these remarks:

Theorem 15.2

Let (M, g_{ab}) be a spacetime with M noncompact. Then every spinor structure on this spacetime arises from some cross-section of the bundle B, as described above. Furthermore, two spinor structures are equivalent if and only if the cross-sections are homotopic, as described above.

It follows from the theorem, in particular, that if a spacetime admits any spinor structure at all, then it always admits at least four – namely, those that arise from the given one by the various combinations of time- and space-reversal. Consider, for example, Minkowski spacetime. Then there is a natural cross-section of B for this spacetime, constructed as follows. Map all tangent spaces of M to each other via parallel transport, and then map the tangent spaces, so identified, to (T', g'). This cross-section produces the standard spinor structure on Minkowski spacetime. We obtain three others by applying to this one the various time- and space-reversals. These four spinor structures exhaust the possibilities for Minkowski spacetime. This follows from the fact that two cross sections of the bundle B are homotopic to each other if and only if they manifest the same orientations.

Consider, to take a second example, Minkowski spacetime with a timelike 2-plane removed. Let, in this spacetime, γ be a closed curve that "goes once around the removed 2-plane". This γ cannot be contracted to a point in M. Indeed, the first homotopy group of M is Z, the additive group of integers. (An integer n in Z corresponds to a curve that traverses γ n times.) Denote by ϕ the standard angular function on M, so, on traversing γ, ϕ goes from 0 to 2π. Now fix a time- and space-orientation on this spacetime. Then there are precisely two spinor structures compatible with these orientations. One is the standard one that comes from the cross-section, C, that arises via parallel transport in this flat spacetime. The other is constructed from this one, as follows. For each point p of M, first map the tangent space T_p at p to T' by the cross-section C, but then apply a spatial rotation in T' through angle $\phi(p)$. Thus, on going around the curve γ, the map to T' undergoes a rotation of 2π relative to the parallel-transported map. This cross-section, C' is not homotopic to C, and so gives rise to a different spinor structure. Other cross-sections can be obtained by applying to C rotations through angle $n\phi(p)$, where n is any integer. But all of these are homotopic to the cross-sections we have already listed: to C when n is even; and to C' when n is odd. (This follows from the fact that a curve in the Lorentz group corresponding to a $2\pi n$ rotation is homotopic to a point if and only if n is even.) Thus, there is a total of eight spinor structures on this spacetime, two for each set of orientation-choices. Do these spinor structures produce, via the Dirac equation, physically different electrons?

It turns out that the general situation is similar to that of the above example. Fix a spacetime (M, g_{ab}), together with some cross-section C of the bundle B. Then all the cross-sections, compatible with the orientations provided by C, are classified, up to homotopy, by homomorphisms from the first homotopy group of M, $\pi_1(M)$, to Z_2, the multiplicative group of integers $(+1, -1)$. In other words, to construct a new spinor structure, one must first specify a map that assigns to each closed curve in M passing through some fixed point $p \in M$, one of $(+1, -1)$, where this specification is such that i) curves that can be continuously deformed to each other are assigned the same integer, and ii) the closed curve that results from first traversing one closed curve and then another is assigned the integer given by the product of the integers assigned to the original curves. Space-times with complicated connectivity can have large numbers of inequivalent spinor structures.

15.5 Lie and Other Derivatives

Fix a manifold M. Recall how Lie derivatives work. To each tangent vector field ξ^a on M, we assign an operator, $\mathcal{L}_\xi$, which acts on every tensor field $T^{\cdot\cdot}{}_{\cdot\cdot}$, of arbitrary rank, on M, returning a tensor field, $\mathcal{L}_\xi T^{\cdot\cdot}{}_{\cdot\cdot}$, of the same index structure. There is one and only one such assignment having the following three properties: i) Each operator $\mathcal{L}_\xi$ is additive, satisfies the Leibnitz rule, and commutes with contraction; ii) applied to any scalar field, each $\mathcal{L}_\xi$ is the ξ-directional derivative; and iii) the operators $\mathcal{L}_\xi$ satisfy the commutation relation $\mathcal{L}_\xi\mathcal{L}_\eta - \mathcal{L}_\eta\mathcal{L}_\xi = \mathcal{L}_{(\mathcal{L}_\xi\eta)}$ (which is really just a variant of the Leibnitz rule). The operator $\mathcal{L}$ thus characterized is, of course, the Lie derivative. It can be defined in at least three different, but equivalent, ways: i) Introduce any derivative operator, ∇_a, on M, write down a formula for the action of $\mathcal{L}_\xi$ in terms of that ∇_a, and note that the result is independent of the choice of derivative operator. ii) Introduce the one-parameter family of local diffeomorphisms on M generated by ξ^a, apply those diffeomorphisms to the tensor field $T^{\cdot\cdot}{}_{\cdot\cdot}$ to obtain a one-parameter family of tensor fields on M, take the parameter-derivative of this family, and evaluate at parameter-value zero. iii) Define the action of $\mathcal{L}_\xi$ on tangent vector fields using the commutation relation, and then extend the action to all tensor fields using property i) above.

Next, fix a spacetime with spinor structure. It is natural to ask whether the action of the Lie derivative, $\mathcal{L}_\xi$, on tensor fields on M, can be extended to include also spinor fields on this spacetime. That is, we ask, for each tangent vector field ξ^a, for an operator, also written $\mathcal{L}_\xi$, on spinor fields satisfying conditions i)–iii) above and also having the property that, when applied to the spinor-representation of a tensor field, it reproduces the original Lie derivative of that tensor field. This last condition can be written as $\mathcal{L}_\xi\sigma^a{}_{BB'} = 0$. It is easy to show that there exists no such extension. Indeed, if there were one, then it would follow from the Leibnitz rule and antisymmetry of $\mathcal{L}_\xi\epsilon_{AB}$ that $\mathcal{L}_\xi(\epsilon_{AB}\bar{\epsilon}_{A'B'})$ is a multiple of $\epsilon_{AB}\bar{\epsilon}_{A'B'}$. In other words (since $\epsilon_{AB}\bar{\epsilon}_{A'B'}$ is the spinor representation of the metric), it would follow that ξ^a is a conformal Killing field. But there always exist tangent vector fields ξ^a on (M, g_{ab}) that are not conformal Killing fields.

Thus, there is in general no such thing as "the Lie derivative of a spinor field". This is not surprising, for spinor fields are "linked to the light cones", and we would not expect them to be Lie-derivable by fields ξ that do not respect those light cones. One might imagine that it would be possible to invent weaker conditions for what constitutes a "Lie derivative", such that, under these weakened conditions, there *is* a unique natural operator $\mathcal{L}_\xi$ on spinor fields. But this program has not turned out to be fruitful.

The remarks above suggest that it should be possible to take Lie derivatives of spinor fields by vector fields ξ^a that "respect the light cones", in the sense that they are conformal Killing vectors. This is indeed the case. For ξ^a a conformal Killing field, we have

$$\nabla_a\xi_b = \phi_{AB}\bar{\epsilon}_{A'B'} + \bar{\phi}_{A'B'}\epsilon_{AB} + 4\kappa\epsilon_{AB}\bar{\epsilon}_{A'B'}\,, \qquad (15.15)$$

for some unique symmetric spinor field ϕ_{AB} and real scalar field κ. (This equation is precisely the assertion that $\nabla_a\xi_b$ differs from $4\kappa g_{ab}$ by some skew tensor field.) We now define an operator $\mathcal{L}_\xi$ on spinor fields as follows. First, set

$$\mathcal{L}_\xi\alpha_A = \xi^m\nabla_m\alpha_A + \phi_A{}^M\alpha_M + \kappa\alpha_A\,, \qquad (15.16)$$

$$\mathcal{L}_\xi\beta^A = \xi^m\nabla_m\beta^A - \phi_M{}^A\beta^M - \kappa\beta^A\,. \qquad (15.17)$$

Now define the action of $\mathcal{L}_\xi$ on one-index primed spinor fields by taking the complex conjugates of these formulae; and on multi-index spinor fields by using the Leibnitz rule. There results an operator, $\mathcal{L}_\xi$, on spinor fields of arbitrary index structure. It is easy to check that this family of operators (one for each conformal Killing field ξ) has all the required properties. (Note that property iii) makes sense, for the Lie bracket of two conformal Killing fields is a conformal Killing field.) In particular, every Killing field on (M, g_{ab}) is also a conformal Killing field (the case $\kappa = 0$ above); and thus Lie derivatives of spinor fields by Killing vector fields make sense.

One might hope that the Lie-derivative operation defined above (for conformal Killing fields ξ) is the unique one having the properties there listed. In fact, it is not. Indeed, fix any real scalar field ρ on M, and introduce, on the right-hand sides of the two equations above, the terms $\mathrm{i}\mathcal{L}_\xi\rho$ and $-\mathrm{i}\mathcal{L}_\xi\rho$, respectively. The result again satisfies all the conditions for a Lie derivative. This reflects a kind of "gauge freedom" inherent in spinors. Thus, while there is a variety of operators $\mathcal{L}_\xi$ satisfying the conditions we have listed, there is a "natural" choice among these, namely the operator given by (15.16) and (15.17).

The Lie derivative of a geometrical object describes how that object changes under "diffeomorphisms differing infinitesimally from the identity". The behavior of

spinors under full diffeomorphism is exactly what one would expect. Fix a spacetime, (M, g_{ab}). Then each diffeomorphism ψ on the manifold M defines a mapping that sends each tensor field $\xi^{a..c}{}_{b..d}$ on M to a new tensor field, $\psi^*\xi$, on this manifold. In particular, ψ sends the metric g_{ab} to some new metric, $g' = \psi^* g$, on M. Now let there be given a spinor structure on (M, g_{ab}). Then this diffeomorphism ψ sends this spinor structure to a spinor structure on (M, g'_{ab}). Furthermore, ψ sends each g-spinor field on M to a g'-spinor field. There is, of course, no natural way to turn this g'-spinor field back into a g-spinor field.

We remark that a similar situation obtains for derivative operators. Recall that a derivative operator, ∇_a, on a general manifold M can be applied to any tensor field $T^{\cdot\cdot}{}_{\cdot\cdot}$ on M, returning a tensor field, $\nabla_a T^{\cdot\cdot}{}_{\cdot\cdot}$. Now consider a spacetime, with spinor structure. Can every such derivative operator ∇_a on the underlying manifold M be extended to act also on spinor fields, retaining the usual properties (additive, Leibnitz, reducing to the original ∇_a for tensor fields, etc.)? The answer is no. And the argument is the same as that above: If there were such an extension, then it would follow that $\nabla_c(\epsilon_{AB}\bar{\epsilon}_{A'B'})$ is a multiple of $(\epsilon_{AB}\bar{\epsilon}_{A'B'})$, i. e., that $\nabla_c g_{ab}$ is a multiple of g_{ab}. But there always exist derivative operators on M that do not have this property. Thus, not every derivative operator on M can be extended, in a natural way, to act also on spinor fields. But, as the remarks above suggest, there does exist such an extension whenever the original derivative operator has the property that $\nabla_c g_{ab}$ is a multiple of g_{ab}, i. e., has the property that it is "light-cone preserving". It turns out that, for a derivative operator ∇_a satisfying this condition, there exists a variety of extensions of ∇_a to spinor fields. But, just as was the case for the Lie derivative, there is singled out, from among all these extensions, one natural one.

There is, of course, always one derivative operator ∇_a satisfying that $\nabla_c g_{ab}$ is a multiple of g_{ab}, namely that satisfying $\nabla_c g_{ab} = 0$. Other such derivative operators are those associated with conformally related metrics, $\Omega^2 g_{ab}$. There are still others.

15.6 4-Spinors

In particle physics, it is common to use what are called "4-spinors". These are closely related to the spinors we have discussed in Sect. 15.1; and in this section we briefly summarize what that relationship is.

Roughly speaking, a "4-spinor" is a pair of "2-spinors", one with an unprimed and one with a primed index: $(\xi^A, \eta_{A'})$. These pairs clearly form a complex, four-dimensional vector space. There is a complex-conjugation operation, which sends this vector to $(\bar{\eta}^A, \bar{\xi}_{A'})$. The "Dirac matrices" play the role of the soldering form: They connect these spinors to vectors. The "derivative" on 4-spinor fields arises directly from the natural derivative operator, ∇_a, on 2-spinors. Thus, for instance, the Dirac equation (15.10) can be written as a first-order, linear equation on a single 4-spinor field.

The paragraph above summarizes the structure of 4-spinors. But it is incomplete – particularly with respect to issues of orientations. For example, the mere existence of 2-spinors on a spacetime imposes on that spacetime a preferred choice of time- and space-orientations, while 4-spinors do not. We discuss below how to account correctly for the orientations.

A *4-spinor space* consists of i) a pair of complex, two-dimensional vector spaces, each with a preferred nonzero antisymmetric spinor; and ii) a pair of antilinear isomorphisms between the vector spaces, differing only by sign, each of which sends the preferred antisymmetric spinor of one space to that of the other.

Each of the vector spaces specified in i), then, has the structure of a spinor space. But this pair of vector spaces is to be "unordered", i. e., neither one is singled out as the "primary" vector space, the other one being constructed from it. The antilinear isomorphisms specified in ii) are "complex-conjugation operations". They take elements of one of the vector spaces to those of the other. Condition ii) specifies there are to be given two candidates for complex conjugation, differing by sign. Again, these two candidates are to be unordered. Thus, in essence, in a 4-spinor space we do not know which spinors have been assigned the privilege of having the unprimed indices and which must make do with the primed; and, while we do have complex conjugation of one-index spinors, we have that operation only up to sign. Note that this sign ambiguity for complex conjugation disappears when applied to spinors with an even number of indices. Thus, the demand that the two preferred antisymmetric spinors be complex conjugates of each other is meaningful. Note that a 4-spinor space

has the structure of a complex, four-dimensional vector space.

We may construct a 4-spinor space as follows. Fix a regular spinor space (V, ϵ_{AB}). Let the vector spaces be V together with its complex conjugate $\overline{V}$, and let the antisymmetric spinors be ϵ_{AB} and $\overline{\epsilon}_{A'B'}$. Let the two antilinear isomorphisms consist of the usual complex-conjugation operation together with that operation with a sign reversal. Finally, unorder, i. e., "forget" that we started with V and constructed $\overline{V}$ from it; and that we started with one preferred complex-conjugation operation and constructed the other from it.

Fix a 4-spinor space. Take the tensor product of the two vector spaces. The complex conjugation operates on this tensor-product without sign ambiguity. Denote by T the real, four-dimensional vector space of all elements of this tensor product that are self-adjoint, i. e., that are equal to their complex conjugates. The two alternating tensors define, just as for 2-spinors, a Lorentz-signature metric g on this T.

Thus, from any 4-spinor space we construct a Lorentz vector space. Two things should be noted regarding this construction. First, this is genuinely a construction from a 4-spinor space, i. e., it does not require that there be singled out one specific choice from the two complex vector spaces, nor one specific choice from the two complex-conjugation operations. And second, the Lorentz vector space (T, g) that arises from this construction comes with no preferred time- or space- or total-orientation.

Suppose that we select one of the two antilinear isomorphisms, and designate it as the official "overline-operation". Then this selection induces a time-orientation on (T, g), namely that for which the future light cone contains elements given by the tensor product of one element of one of the vector spaces with the overline of that element. Had there been chosen the other complex-conjugation operation, then the opposite time-orientation would have been induced.

Next, select one of the two vector spaces as the "primary" one. That is, choose one of the vector spaces to have its elements denoted by unprimed indices, the other by primed. Then this selection assigns a particular total-orientation to T, namely that associated with the alternating tensor ϵ_{abcd} given by (15.5). Note that if you reverse this choice, i. e., reverse the roles of primed and unprimed indices on the right-hand side in (15.5), then you reverse the sign of this alternating tensor. In other words, you reverse the assigned total orientation.

To summarize, a 4-spinor space represents, effectively, "2-spinors, but stripped of the structure that singles out orientations". One can then restore that additional structure by hand, and thereby recover those orientations.

Next, let (M, g) be a spacetime. A *4-spinor structure* on (M, g) is a 4-spinor bundle over M together with a soldering form, which provides an isometry, at each point, between the above-constructed Lorentz vector space and the tangent space at that point. Note that a spacetime can have a 4-spinor structure without being either time-orientable, or space-orientable, or total-orientable. Indeed, fix, in a spacetime with 4-spinor structure, any closed curve γ in M. Then, for example, the result of traversing that curve in the spacetime will reverse time-orientation if and only if traversing that curve reverses the sign of the complex-conjugation operation. Similarly for the other orientations. Thus, failure of a spacetime to be time- and/or space-orientable is no barrier to its having 4-spinors. However, the additional topological condition, discussed in Sect. 15.4, for a spacetime to have 2-spinors is also required in order to have 4-spinors.

Finally, we introduce a derivative operator on 4-spinor fields, in exactly the same manner as for 2-spinors. We are thus able to write out systems of differential equations on such fields. Specifically, we may write out, for example, the Dirac equation (15.10). This now becomes a single, linear, first-order equation on a single 4-spinor field. We conclude from all this, among other things, the following: Electrons make sense in a spacetime, even if it fails to be time- or space-orientable.

15.7 Euclidean Spinors

Fix a complex, two-dimensional vector space V. Then, just as in Sect. 15.1, we may apply dualization and complex conjugation to V, and then take tensor products between the resulting vector spaces. There results spinors – objects with primed and unprimed subscripts and superscripts – over V. The next step, in Sect. 15.1, was to introduce, as a fundamental object, an antisymmetric spinor ϵ_{AB}. This spinor was then incorporated

into the notation by using it to raise and lower indices; and once so incorporated ϵ practically disappeared from the formalism.

This time, we proceed in a slightly different way. Instead of just ϵ_{AB}, we introduce *two* fundamental objects. One is the same nonzero, antisymmetric spinor ϵ_{AB} that we had before. The other is a spinor $t_{AA'}$, which satisfies the following three conditions: It is self-adjoint (i.e., satisfies $\bar{t}_{A'A} = t_{AA'}$); it is positive-definite (i.e., is such that, for any nonzero α^A, $t_{AA'}\alpha^A\overline{\alpha}^{A'} > 0$); and it is normalized with respect to ϵ by

$$t_{A[A'}t_{|B|B']} = \frac{1}{2}\epsilon_{AB}\overline{\epsilon}_{A'B'} \,. \tag{15.18}$$

This $t_{AA'}$ defines a positive-definite, Hermitian inner product on V: $\langle\alpha|\beta\rangle = \alpha^A\overline{\beta}^{A'}t_{AA'}$. Thus, this vector space V, with this inner product, has the structure of a Hilbert space. But it has more structure than this: There is also specified the ϵ_{AB}.

We next incorporate these two objects into the formalism, as follows. We incorporate ϵ_{AB}, just as before, by using it, together with its inverse, complex conjugate, and inverse complex conjugate, to raise and lower spinor indices. We incorporate $t_{AA'}$ by using it to convert primed indices to unprimed. To this end we define an adjoint operation, on spinors with only unprimed indices, as follows. For $\alpha^{A\cdots C}{}_{B\cdots D}$ any such spinor, we set

$$\alpha^{\dagger A\cdots C}{}_{B\cdots D} = (-1)^s t^A{}_{A'}\cdots t^C{}_{C'} \times t_B{}^{B'}\cdots t_D{}^{D'}\overline{\alpha}^{A'\cdots C'}{}_{B'\cdots D'} \,, \tag{15.19}$$

where s is the number of superscripts of α. Thus, the adjoint of a scalar is its complex conjugate, while $\epsilon^{\dagger}{}_{AB} = \epsilon_{AB}$. Taking the adjoint commutes with outer product and contraction, and so with the raising and lowering of indices. We have, for any spinor α, $\alpha^{\dagger\dagger} = (-1)^r\alpha$, where r is the total number of indices of α.

We may now deal exclusively with spinors with unprimed indices, the complex-conjugation operation having been replaced in favor of this adjoint. With this convention, the spinor $t_{AA'}$ (now buried in the adjoint) disappears entirely from the formalism. Call a spinor *self-adjoint* if it is equal to its adjoint. Then every spinor with an even number of indices can be written uniquely as the sum of a self-adjoint and an anti-self-adjoint spinor. But for spinors with an odd number of indices the situation is completely different: If a spinor with an odd number of indices is self-adjoint or anti-self-adjoint, then it vanishes. Note that quite generally $\alpha^{\dagger A\cdots D}\alpha_{A\cdots D}$ is nonnegative, vanishing only when $\alpha_{A\cdots C}$ itself vanishes.

Fix $(V, \epsilon_{AB}, t_{AA'})$, as described above. Consider the vector space K of all spinors ξ_{AB} that are symmetric and self-adjoint: $\xi_{AB} = \xi^{\dagger}{}_{(AB)}$. This K is a real three-dimensional vector space. On this vector space, we can form the inner product, $\langle\phi, \tau\rangle = \phi^{AB}\tau_{AB}$. Note that this inner product is real, by self-adjointness, and positive-definite, by positive-definiteness of $t_{AA'}$. In short, this vector space K, with this inner product $\langle\ ,\ \rangle$, is a Euclidean vector space.

Now fix any Euclidean vector space, i.e., any real three-dimensional vector space W with positive definite metric q_{ab}. By a *spinor structure* on (W, q_{ab}), we mean a complex two-dimensional vector space V, with spinors ϵ_{AB} and $t_{AA'}$, as described above, together with an isometry between the Euclidean vector space K, $\langle\ ,\ \rangle$ constructed from $(V, \epsilon_{AB}, t_{AA'})$ and (W, q_{ab}). We may describe this isometry by means of a soldering form, $\sigma^a{}_{BC}$.

Thus, the construction of spinors for a Euclidean vector space is completely analogous to the construction for a Lorentz vector space. The key difference is that, for the Euclidean case, we introduce one additional fundamental object, the spinor $t_{AA'}$. (Think of this $t_{AA'}$ as a "fixed timelike vector" in the Lorentz vector space, and of the Euclidean vector space as constructed of vectors that are orthogonal to $t_{AA'}$.) Just as we used Lorentz spinors and their Lorentz vector space to install spinors on spacetimes, we now use Euclidean spinors and their Euclidean vector space to install spinors on certain Riemannian manifolds.

Fix a Riemannian 3-manifold, (N, q_{ab}), so q_{ab} is a positive-definite metric on the manifold N. Then at each point of N, the tangent space has the structure of a Euclidean vector space. By a *spinor structure* on (N, q_{ab}) we mean a smooth assignment of a spinor structure to the tangent space of each point of N. Applying the construction above to each point of N, we acquire spinor fields (all with unprimed indices) on N. On those fields we have the usual operations: addition, outer product, contraction and the adjoint operation. Just as in the spacetime case, we have a derivative operator on these spinor fields: There exists one and only one extension of the derivative operator D_a of (N, q_{ab}) to spinor fields that i) is additive, commutes with contraction and the adjoint operation, and satisfies the Leibniz rule under outer product; and ii) annihilates ϵ_{AB} and the soldering form.

An important application of spinor fields on a Riemannian 3-manifold is to Witten's proof [15.9] that the

total mass, measured at spatial infinity, of certain spacetimes must be positive. Consider an initial-data set, (N, q_{ab}, p_{ab}). Here, N is a 3-manifold, q_{ab} a positive-definite metric (the induced metric) on N, and p_{ab} a symmetric tensor field (the extrinsic curvature) on N. Let the 3-manifold N be diffeomorphic with R^3. Let this initial data set be asymptotically flat in a suitable sense, i. e., let q_{ab} approach a flat metric and p_{ab} approach zero at infinity, at appropriate rates. It is known that, under these conditions, there exists a spinor field λ^A on N that satisfies the Witten equation,

$$D_{AB}\lambda^B = \frac{-\mathrm{i}}{\sqrt{8}} p^m{}_m \lambda_A \,, \tag{15.20}$$

and approaches a constant spinor at infinity, in an appropriate sense. In fact, there exists one and only one such solution, for each value of the "asymptotic constant". Fix such a solution λ^A. We then write down a certain integral, whose integrand is quadratic in λ^A and its derivatives. This integral has two properties. First, the integrand is nonnegative, provided only that the sources for the original initial data set satisfy an energy condition. It follows that the value of integral itself is nonnegative. Second, the integrand is an exact divergence, and so, using Gauss' law, we may rewrite that integral as a surface integral. It then turns out that this surface integral is precisely a certain component of the total mass-momentum of the initial data set, namely the component along the null direction determined by the asymptotic behavior of the solution λ^A. These two properties imply, then, that a certain null component of the total mass-momentum 4-vector is nonnegative. Furthermore, we are guaranteed that there are enough solutions of (15.20) that we can test in this way every null component of that total mass-momentum. It follows that the asymptotic 4-vector representing the total mass-momentum must be future-directed timelike or null. This is the Witten proof of the positive-mass theorem. For details, see Chap. 18.

What is striking about this proof is that it requires, in an essential way, the use of spinors.

Closely related to Euclidean spinors are the spin-weighted functions on the 2-sphere [15.10]. Fix a Euclidean spinor space $(V, \epsilon_{AB}, t_{AA'})$. Denote by S the set of all unit vectors in the corresponding Euclidean vector space, i. e., the set of symmetric, self-adjoint spinors satisfying $\phi^{AB}\phi_{AB} = 1$. Then S has the structure of a 2-sphere. We next introduce certain complex-valued functions on this 2-sphere. Let $\tau_{A\cdots D}$ be any totally symmetric spinor, with $2l$ indices, where l is a nonnegative integer. The function ψ on S given by $\psi(\phi^{AB}) = \tau_{A\cdots D}\phi^{AB}\cdots\phi^{CD}$ is call a *spherical harmonic* of rank l. The real and imaginary parts of this ψ are associated with the self-adjoint and anti-self-adjoint parts of $\tau_{A\cdots D}$, respectively. Note that, had we not imposed symmetry on the $\tau_{A\cdots D}$ above, then we could still have defined the function ψ as above, but now it would be equal to a sum of spherical harmonics. This follows from the decomposition of $\tau_{A\cdots D}$ in terms of totally symmetric spinors. The collection of all spherical harmonics, of fixed rank l, forms a vector space, of (complex) dimension $(2l+1)$. The standard formula expressing the product of two spherical harmonics as a linear combination of spherical harmonics is recovered, in spinor language, from the formula expressing the symmetrized product of two symmetric spinors in terms of totally symmetric spinors. It follows immediately from this characterization that the ranks of the spherical harmonics in the product range from $l+l'$ to $|l-l'|$, where l and l' are the ranks of the factors.

We next claim that every spinor $\phi^{AB} \in S$ can be written in the form $\phi^{AB} = \mathrm{i}\sqrt{2}\alpha^{\dagger(A}\alpha^{B)}$, where α^A is some unit spinor: $\alpha^{\dagger A}\alpha_A = 1$. To see this, note that, as in Sect. 15.1, we may write any symmetric spinor ϕ^{AB} in the form $\phi^{AB} = \alpha^{(A}\beta^{B)}$, for some α, β. But self-adjointness of ϕ^{AB} implies that α and β must, up to a factor, be adjoints of each other. The factor $\mathrm{i}\sqrt{2}$ gets the normalization right.

Note that, in the decomposition above, $\phi^{AB} \in S$ determines α^A uniquely up to phase. It turns out that this phase has a simple geometrical interpretation: It represents a direction in the 2-sphere S at the point $\mathrm{i}\sqrt{2}\alpha^{\dagger(A}\alpha^{B)} \in S$. To see this, fix unit α^A, and consider $\mathrm{i}(\alpha^A\alpha^B - \alpha^{\dagger A}\alpha^{\dagger B})$. This spinor is symmetric and self-adjoint, and is orthogonal to ϕ^{AB}. Thus, $\mathrm{i}(\alpha^A\alpha^B - \alpha^{\dagger A}\alpha^{\dagger B})$ defines a first-order change in ϕ^{AB} that preserves, to first order, its symmetry, self-adjointness, and uniticity. That is, it defines a direction in the 2-sphere, S, of symmetric, self-adjoint, unit ϕ^{AB}. Multiplying α^A by $e^{\mathrm{i}\lambda}$ rotates this direction through angle 2λ. So, e.g., α^A and $-\alpha^A$ define the same direction. We conclude, then, that the manifold of all unit spinors α^A is a double covering of the direction bundle of the 2-sphere.

A complex-valued function f on this manifold of unit spinors is called *spin weighted* provided it has the following property: For any unit α^A and real number λ, $f(e^{\mathrm{i}\lambda}\alpha^A) = e^{2\mathrm{i}s\lambda}f(\alpha^A)$. Here, s is a half-integer (positive or negative), called the *spin weight* of f. The spin-weighted functions, of given weight, form a complex vector space (of infinite dimension). The product

of two spin-weighted functions is a spin-weighted function whose weight is the sum of those of the factors. Complex conjugation reverses the sign of spin weight. The functions of spin-weight zero (i. e., that are independent of the phase of α) are just ordinary complex-valued functions on the 2-sphere S.

A simple class of spin-weighted functions consists of those that are "polynomial". Fix a totally symmetric spinor $\tau_{A\cdots D}$ (not necessarily with an even number of indices). Consider the function given by $f(\alpha^A) = \tau_{A\cdots D}\alpha^A\cdots\alpha^B\alpha^{\dagger C}\cdots\alpha^{\dagger D}$, i. e., the result of contracting τ with some number a of α's, and some number b of $\alpha^\dagger$'s. This is clearly a spin-weighted function, of spin-weight $(a-b)/2$. These functions are called the *spin-weighted spherical harmonics*. Again, had we not imposed symmetry on $\tau_{A\cdots D}$ above, then we could still have defined the function f as above, but now it would be equal to a sum of spin-weighted spherical harmonics. This follows, decomposing τ in terms of totally symmetric spinors. In the case $a = b$, i. e., spin-weight zero, the spin-weighted spherical harmonics reduce to the ordinary spherical harmonics, introduced above. By the *rank*, l, of a spin-weighted spherical harmonic, we mean half the number of indices on $\tau_{A\cdots D}$, so l is a nonnegative half-integer, and $-l \le s \le l$. The spin-weighted spherical harmonics of given spin weight and rank form a complex vector space of dimension $(2l+1)$. Indeed, it is clear from the definition that this vector space is precisely that of totally symmetric, $2l$-index spinors. In other words, fixing the rank l once and for all, the spin-weighted functions for various choices of spin-weight s merely constitute different ways of talking about the same thing: Totally symmetric spinors with $2l$ indices. (Usually, it is easier just to stick with the spinors.) One can now derive directly the Clebsch–Gordon formulae, which express a product of spin-weighted spherical harmonics as a linear combination of spin-weighted spherical harmonics. If you are adept at manipulating spinors, then you are automatically adept at manipulating spin-weighted spherical harmonics.

It is convenient to be able to take derivatives[1] of spin-weighted functions. To this end, we define an operator, $\eth$, with action $\eth f = \alpha^A \partial f/\partial\alpha^{\dagger A}$. Thus, the action of $\eth$ on a spin-weighted function raises the spin weight by 1. Applied to a spin-weighed spherical harmonic, $\eth$, returns another spin-weighted spherical harmonic, having the same rank (indeed, having the same underlying spinor $\tau_{A\cdots D}$), but with different weight. There is a corresponding spin-weight-lowering operator $\bar{\eth} f = \alpha^{\dagger A}\partial f/\partial\alpha^A = (\overline{\eth\bar{f}})$. It follows directly from the definition that the commutator of these two has the following action: $[\eth, \bar{\eth}]f = 2sf$, where f is any spin-weighted function and s is its spin weight. Applied to functions of spin-weight zero, the operator $\eth\bar{\eth}$ (or, what, for such functions, is the same thing, $\bar{\eth}\eth$) is equal to minus the Laplace operator on the sphere. The easiest way to prove this is to check that it holds for spherical harmonics, and then use completeness.

It turns out that conformal transformations on the 2-sphere S are generated by changing the choice of the fundamental spinor $t_{AA'}$.

15.8 Bases; Spin Coefficients

Let (T, g_{ab}) be a Lorentz vector space, with spinor structure $(V, \epsilon_{AB}, \sigma^b{}_{AA'})$. By a (normalized) *basis* for V, we mean a pair of spinors, o^A, ι^A, normalized by $o^A\iota_A = 1$. It follows that these two span V. In fact, the general element α^A of V can be written as the linear combination $\alpha^A = (\alpha^M\iota_M)o^A - (\alpha^M o_M)\iota^A$. We obtain from this basis for V also a basis for $\overline{V}$, consisting of $\bar{o}^{A'}, \bar{\iota}^{A'}$. Then any spinor, of any rank, can be expressed in terms of complex numbers, the components of that spinor in these bases.

Fix a normalized basis, o^A, ι^A for V. It is convenient to introduce the following three vectors: $l^a = o^A\bar{o}^{A'}$, $n^a = \iota^A\bar{\iota}^{A'}$, and $m^a = \sqrt{2}o^A\bar{\iota}^{A'}$. The first two are real, the third complex. From the normalization of the V-basis, we find: $l^a l_a = n^a n_a = 0$, $l^a n_a = 1$, $l^a m_a = n^a m_a = 0$, $m^a m_a = 0$, and $m^a\overline{m}_a = -2$. In other words, l^a and n^a are null vectors and the real and imaginary parts of m^a are unit, orthogonal spacelike vectors lying in the 2-plane orthogonal to l^a and n^a. Such a system, $l^a, n^a, m^a, \overline{m}^a$, is called a *null tetrad*. Every complex vector over T can be written as a unique linear combination (with complex coefficients) of these four vectors.

Next, fix, on a spacetime with spinor structure, a basis-field, o^A, ι^A. We introduce the components of the derivatives of these basis spinors. These are organized as follows. First consider the three (complex) vectors $o_A\nabla_b o^A$, $o_A\nabla_b\iota^A$, and $\iota_A\nabla_b o^A$. These give all the information contained in the derivatives of o^A and ι^A. (The fourth combination, $o_A\nabla_b\iota^A$, is equal to $\iota_A\nabla_b o^A$, as a consequence of the normalization condition, $o^A\iota_A =$

1.) Now contract these three vectors in turn with with each of l^b, n^b, m^b, and $\overline{m}^b$. The 12 complex scalar fields that result, called the *spin coefficients*, carry all the information in the derivatives of o^A and ι^A. Note that the derivatives of the null-tetrad vectors, l^a, n^a, m^a and $\overline{m}^a$, can also be expressed in terms of these spin coefficients. The idea of the spin-coefficient formalism [15.11, 12] is that, since there are only 12 scalar fields, it becomes just feasible to assign a separate Greek letter to each one, and then write out everything explicitly in terms of components.

The next step is to introduce the components of the derivative operator. Set

$$D = l^a \nabla_a \, ; \quad d = n^a \nabla_a \, ;$$
$$\partial = m^a \nabla_a \, ; \quad \overline{\partial} = \overline{m}^a \nabla_a \, . \tag{15.21}$$

These four differential operators carry all the information in ∇_a.

The final step [15.11, 12] is to generate the so-called spin-coefficient equations. This is a system of equations that are linear in the first derivatives of the spin coefficients, but also involve nonlinear terms algebraic in those coefficients. The spin-coefficient equations are divided into two sets.

The first set of equations consists of those involving the curvature tensor. Consider the two equations $\nabla_{[a}\nabla_{b]}(o^C) = -\rho_{ab}{}^C{}_D o^D$ and $\nabla_{[a}\nabla_{b]}(\iota^C) = -\rho_{ab}{}^C{}_D \iota^D$, where $\rho_{ab}{}^C{}_D$ is given in terms of the curvature by (15.12). Take the components of these equations, using l^a, n^a, m^a, and $\overline{m}^a$ for tensor indices, and o^A, ι^A for spinor indices. The right-hand sides become components of the curvature tensor; while the left-hand sides, on differentiating by parts, become expressions linear in the derivatives of the spin coefficients. In this way, we generate equations expressing the curvature tensor in terms of the spin coefficients and their derivatives.

The second set of equations does not involve the curvature tensor. Consider, e.g., the identity $\nabla_{[a}\nabla_{b]}(o^C \iota_C) = 0$. Again, take components of this equation using the null tetrad; expand the left-hand side; and, again, differentiate the left-hand side by parts. We thus generate a set of equations involving only spin coefficients and their derivatives.

Equations for other fields are also reduced to components. For Maxwell's equations, for example, we first introduce the three complex scalar fields given by the components, in our spinor basis, of the Maxwell spinor, ϕ_{AB}. The components of Maxwell's equations then become a system of equations involving the spin coefficients and these Maxwell components, linear in the derivatives of the latter.

The spin-coefficient formalism is generally most useful when one can choose the spinor basis such that a number of the spin coefficients vanish. Typically, this occurs when the spacetime has one or two preferred null vector fields, e.g., when it is algebraically special.

15.9 Variations Involving Spinors

In this section, we discuss how to set up variational problems when spinors are involved. It turns out that there are a few subtle issues.

Consider, as an example, the Dirac equation, (15.10). A Lagrangian for this system is given by

$$\begin{aligned} L = \int \Big[&\mathrm{i}\left(\xi^A \nabla_{AA'} \overline{\xi}^{A'} - \overline{\xi}^{A'} \nabla_{AA'} \xi^A\right) \\ &- \mathrm{i}\left(\eta^{A'} \nabla_{AA'} \overline{\eta}^A - \overline{\eta}^A \nabla_{AA'} \eta^{A'}\right) \\ &+ \mathrm{i}\sqrt{2} m \left(\overline{\xi}^{A'} \eta_{A'} - \xi^A \overline{\eta}_A\right) \Big] \epsilon_{cdef} \, \mathrm{d}V^{cdef} \, . \end{aligned} \tag{15.22}$$

What this means is the following. Fix the spacetime (M, g_{ab}), once and for all. Then given any pair of spinor fields $(\xi^A, \eta_{A'})$, we may compute L via (15.22). We now ask for a such a pair having the following property: On varying $(\xi^A, \eta_{A'})$ the L given by (15.22) does not change to first order. This requirement makes sense, because we know what it means to change the spinor fields, $(\xi^A, \eta_{A'})$, on a fixed background spacetime. The calculation is the standard one, and the result is what we expect: This requirement is precisely the requirement that $(\xi^A, \eta_{A'})$ satisfies the Dirac equation (15.10).

Next, let us try to use this Lagrangian to compute the stress-energy of the Dirac field. To do this, we must vary the metric g_{ab} in the Lagrangian (15.22), keeping the spinor fields $(\xi^A, \eta_{A'})$ fixed. But what does this mean? Spinors owe their very existence to the metric. The spinors associated with one metric are completely different objects from those associated with another. So how can one maintain the "same" spinors, while varying the metric? Since the spinor manifestation of the metric is the alternating spinor ϵ_{AB}, one might try

the following strategy. Write the Lagrangian entirely in terms of spinors, making explicit each appearance of ϵ_{AB}. Then vary this ϵ_{AB}, keeping the other spinors fixed. Unfortunately, this does not work. The variation of ϵ_{AB} must be a complex multiple of ϵ_{AB}, and so this variation is described by a single complex function. But surely this is no substitute for the ten components required for the variation of g_{ab}. Does there exist, then, any sensible way at all to extract a stress-energy from (15.22)?

It turns out that there does. The idea is to fix, once and for all, the spinor bundle with its alternating spinor, ϵ_{AB}. We then "vary the metric g_{ab}" indirectly, by varying the soldering form $\sigma^a{}_{BB'}$. Indeed, from (15.1), we see that in order to generate a variation $h_{ab} = \delta g_{ab}$ in the metric, we must make a variation in the soldering form given by

$$\delta\sigma^a{}_{BB'} = -\frac{1}{2}h^a{}_{BB'} \,. \tag{15.23}$$

Note that the this variation is a symmetric tensor. A variation in $\sigma^a{}_{BB'}$ by an antisymmetric tensor leaves the spacetime metric g_{ab} unchanged. It represents, effectively, an internal change in the spinor structure. So, since we are only interested in metric variations, we consider only those σ-variations given by (15.23). We must next determine the corresponding variation in the derivative operator ∇_a. For its action on tensor fields, $\delta\nabla_a$ is given by the standard formula: For a fixed contravariant vector field k^a, for example, we have

$$\delta\left(\nabla_a k^b\right) = \frac{1}{2}\left(\nabla^b h_{am} - 2\nabla_{(m}h_{a)}{}^b\right)k^m \,. \tag{15.24}$$

There is a similar formula for a general tensor field. In order to determine the variation of ∇_a applied to spinor fields, we proceed as follows. There must (by the properties of a derivative operator) exist a spinor field $L_a{}^M{}_B$ such that, for any fixed spinor field α^A

$$\delta(\nabla_a\alpha^B) = L_a{}^B{}_M\alpha^M \,. \tag{15.25}$$

We must now express this L in terms of h_{ab}. To this end, we first note that $\delta(\nabla_c\sigma^a_{BB'}) = 0$. Expanding the left-hand side, we obtain three types of terms. The first, arising from the variation of σ, is $\nabla_c(\delta\sigma^a{}_{BB'})$, which we evaluate in terms of h_{ab} via (15.23). The second, arising from the variation of ∇_c applied to tensors, we evaluate using (15.24). The third, arising from the variation of ∇_c applied to spinors, we evaluate using (15.25). Equating to zero the sum of all these terms, we obtain the desired expression

$$L_{aBC}\bar{\epsilon}_{B'C'} + \bar{L}_{aB'C'}\epsilon_{BC} = -\nabla_{[b}h_{c]a} \,. \tag{15.26}$$

Note that L_{aBC} is symmetric in indices "B, C", i.e., that, from (15.25), $\delta(\nabla_a\epsilon_{BC}) = 0$. This is what we expect.

So, the general strategy is to write the Lagrangian with all appearances of the soldering form made explicit, and then take the variation of that Lagrangian with respect to that soldering form, using (15.23)–(15.26). Let us now apply this strategy to the Dirac case. For the first step, we have, rewriting (15.22),

$$\begin{aligned} L = \int \Big[&\mathrm{i}\left(\xi^A\nabla_b\bar{\xi}^{A'} - \bar{\xi}^{A'}\nabla_b\xi^A\right)\sigma^b{}_{AA'} \\ &-\mathrm{i}\left(\eta^{A'}\nabla_b\bar{\eta}^A - \bar{\eta}^A\nabla_b\eta^{A'}\right)\sigma^b{}_{AA'} \\ &+\mathrm{i}\sqrt{2}m\left(\bar{\xi}^{A'}\eta_{A'} - \xi^A\bar{\eta}_A\right)\Big]\,\epsilon_{cdef}\,\mathrm{d}V^{cdef} \,. \end{aligned} \tag{15.27}$$

Variation now leads to the stress energy for the Dirac field

$$\begin{aligned} T_{ab} = \mathrm{i}/2\Big(&\xi_A\nabla_b\bar{\xi}_{A'} - \bar{\xi}_{A'}\nabla_b\xi_A + \eta_{A'}\nabla_b\bar{\eta}_A \\ &-\bar{\eta}_A\nabla_b\eta_{A'} + \xi_B\nabla_a\bar{\xi}_{B'} - \bar{\xi}_{B'}\nabla_a\xi_B \\ &+\eta_{B'}\nabla_a\bar{\eta}_B - \bar{\eta}_B\nabla_a\eta_{B'}\Big) \,. \end{aligned} \tag{15.28}$$

Note that there is no "m-term" in this expression for the stress-energy. This is a consequence of the fact that m itself can be eliminated, using (15.10), in favor of derivatives of the Dirac fields. In the spin-zero case, by contrast, no such elimination is available, and thus in that case there do appear terms explicitly involving m. This stress-energy, by virtue of (15.10), is conserved.

[1] Recall a few facts about functions of a complex variable. A complex-valued function f on the complex plane ($z \in C$) is said to be *differentiable* at $z = z_o$ provided: There exist complex numbers σ and τ such that $f(z_o + \delta z) - f(z_o) - \sigma\delta z - \tau\overline{\delta z}$ vanishes to first order in $|\delta z|$. The two complex numbers σ and τ are unique if they exist. They are written as $\mathrm{d}f/\mathrm{d}z$ and $\mathrm{d}f/\mathrm{d}\bar{z}$, respectively. (I find this notation a little confusing, since, after all, f is just a function of $z \in C$.) The function f is said to be *smooth* if all its derivatives, of all orders, exist everywhere and are continuous. Note that smoothness is much weaker than complex-analyticity. Similarly for functions of several complex variables.

References

15.1 F.A.E. Pirani: Spinors, Brandeis Summer Inst. Theor. Phys. 1964 (Prentice Hall, Princeton, NJ 1965)

15.2 R. Penrose: Structure of Spacetime, Battelle Recontres Math. Phys. 1967, ed. by C. DeWitt (W. A. Benjamin, New York 1968)

15.3 R. Penrose: A spinor approach to general relativity, Ann. Phys. **10**, 171–201 (1960)

15.4 P. O'Donnell: *Introduction to 2-Spinors in General Relativity* (World Scientific, New York 2003)

15.5 J.N. Goldberg, R.K. Sachs: A theorem on Petrov types, Acta Phys. Pol. **22**, 13–23 (1962)

15.6 A.Z. Petrov: The classification of spaces defining gravitational fields, Gen. Relativ. Gravit. **32**, 1665–1685 (2000)

15.7 L. Bel: Introduction d'un tenseur du quatrième ordre, C. r. **248**, 1297–1318 (1959)

15.8 R. Geroch: Spinor structure of space-times in general relativity I, J. Math. Phys **9**, 1739–1744 (1968)

15.9 E. Witten: A new proof of the positive energy theorem, Commun. Math. Phys **80**, 381–402 (1981)

15.10 J.N. Goldberg, A.J. Macfarlane, E.T. Newman, F. Rohrlich, E.C.G. Sudarshan: Spin-s spherical harmonics and $\eth$, J. Math. Phys. **8**, 2155–2161 (1967)

15.11 E.T. Newman, R. Penrose: An approach to gravitational radiation by a method of spin-coefficients, J. Math. Phys **3**, 566–768 (1962)

15.12 R. Geroch, A. Held, R. Penrose: A space-time calculus based on pairs of null directions, J. Math. Phys **14**, 874–881 (1973)

16. The Initial Value Problem in General Relativity

James Isenberg

One of the most effective ways of constructing and studying solutions of Einstein's gravitational field equations is via the Initial Value Problem. According to this approach, one constructs spacetime solutions by choosing initial data on a spacelike manifold representing the initial state of a model universe, and one then evolves the data into a spacetime solution representing the full history of that model universe.

A set of initial data cannot be chosen freely: it must satisfy a set of partial differential equations known as the Einstein constraint equations. Not only are these constraint equations a necessary condition on initial data sets; they are as well a sufficient condition for an initial data set to admit evolution into a spacetime solution. After showing how to split the full set of Einstein's field equations into the constraint equations and the evolution equations, we discuss the Well-Posedness Theorem, which shows that indeed all constraint-satisfying data sets can be evolved into spacetime solutions.

Our primary focus is on how to construct and parametrize initial data sets which satisfy the Einstein constraint equations. The Conformal and the Conformal Thin Sandwich Methods both provide ways of turning the constraint equations into a determined nonlinear elliptic system. These equivalent procedures are very effective for initial data sets which involve constant mean curvature or near-constant mean curvature. The challenge is to adapt these methods to more general data sets. An alternative approach for constructing and analyzing solutions of the constraints is via Gluing techniques, which we briefly outline, along with their remarkable applications.

We comment briefly on some of the main questions which arise in studying the long-time behavior of spacetime solutions of Einstein's equations.

16.1 Overview

Ever since Newton's formulation of particle mechanics over three hundred years ago, one of the most widely used methods of modeling physical systems is via an initial value formulation. For particle mechanics, the idea is that one specifies the initial position y_0 and initial momentum p_0 of the particle, and one then determines the particle's path $y(t)$ in time by solving the initial value problem consisting of Newton's equations (an or-

dinary differential equation system)

$$\frac{d}{dt}y(t) = \frac{1}{m}p(t) ,$$
$$\frac{d}{dt}p(t) = F(y,p) ,$$

together with the initial conditions $y(t_0) = y_0$ and $p(t_0) = p_0$. Analogously, to study the vibrational motion $\psi(x,t)$ of a stretched string, one specifies the initial displacement $\psi_0(x)$ and initial velocity $v_0(x)$ of the string, and one then determines the subsequent motion $\psi(x,t)$ by solving the initial value problem consisting of the (first-order) wave equation (a partial differential equation system)

$$\partial_t\psi(x,t) = \nu(x,t) ,$$
$$\partial_t\nu(x,t) = \alpha\partial_{xx}\nu(x,t) ,$$

together with the initial conditions $\psi(x,t_0) = \psi_0(x)$ and $\nu(x,t_0) = \nu_0(x)$.

Although one of the signature properties of Einstein's theory of gravity is its spacetime covariant character, it too admits an initial value formulation. We recall that to model the gravitational interactions of physical systems, general relativity incorporates the physics of the gravitational field into the geometry of spacetime, which is specified by a four-dimensional manifold M^4 together with a metric g of signature $(-,+,+,+)$. Einstein's equations relate the spacetime metric to the nongravitational *matter fields* Ψ which are present in the spacetime by requiring that the Einstein tensor $G_{\mu\nu} := R_{\mu\nu} - (2)Rg_{\mu\nu}$ of g (here $R_{\mu\nu}$ is the Ricci tensor of g and R is its scalar curvature) be proportional to the stress–energy tensor $T_{\mu\nu}$ of Ψ and g; specifically,

$$G_{\mu\nu}[g] = \kappa T_{\mu\nu}[g,\Psi] , \tag{16.1}$$

where $\kappa = 8\pi G_N$ with G_N Newton's gravitational constant. One of the most effective ways to construct and study a solution (M^4, g, Ψ) to Einstein's equations (16.1) is specifying a set of initial data corresponding to an initial *state of the universe*, and then using Einstein's equations (reformulated as an initial value problem) to evolve this data into the corresponding spacetime solution. This chapter discusses the details of this initial value formulation for Einstein's theory: the specific nature of the initial data sets for which the initial value problem works, how one finds such initial data sets, how one reformulates Einstein's equations (16.1) as an initial value problem and verifies that it is well-posed, and some open problems related to the evolution of initial data sets.

One very important feature of the initial value formulation of Einstein's theory is that the initial data must satisfy a set of *constraint equations*. This is a familiar feature of Maxwell's theory of electromagnetism as well: An initial data set for Maxwell's theory consists of a pair of spatial vector fields $(\boldsymbol{E}_0(\boldsymbol{x}), \boldsymbol{B}_0(\boldsymbol{x}))$ which are required to satisfy the Maxwell constraint equations

$$\nabla\cdot\boldsymbol{E}_0 = 0 ,$$
$$\nabla\cdot\boldsymbol{B}_0 = 0 ,$$

(presuming that the charge density is zero). A solution $(\boldsymbol{E}(\boldsymbol{x},t), \boldsymbol{B}(\boldsymbol{x},t))$ is obtained from the initial data via the evolution equations. (Here and throughout this chapter, we set c, the speed of light equal to 1.)

$$\partial_t\boldsymbol{B}(\boldsymbol{x},t) = \nabla\times\boldsymbol{E}(\boldsymbol{x},t) ,$$
$$\partial_t\boldsymbol{E}(\boldsymbol{x},t) = -\nabla\times\boldsymbol{B}(\boldsymbol{x},t) .$$

It is important to note that presuming a data set $(\boldsymbol{E}_0(\boldsymbol{x}), \boldsymbol{B}_0(\boldsymbol{x}))$ satisfies the Maxwell constraints; it follows from the evolution equations that the solution generated from this data set satisfies the constraints $\nabla\cdot\boldsymbol{E}(\boldsymbol{x},t) = 0$ and $\nabla\cdot\boldsymbol{B}(\boldsymbol{x},t) = 0$ for all times t.

To understand the initial value formulation of Einstein's theory and how it produces spacetimes (M^4, g, Ψ) which satisfy Einstein's equations, it is useful to start examining spatial foliations of such a spacetime, with a focus on the geometric fields induced on the spatial leaves of such a foliation and how they relate to the spacetime metric g. We do this in Sect. 16.2. Included in this section is a discussion of the Gauss–Codazzi–Mainardi equations which relate the spacetime curvature of g to the induced fields. Based on these equations, we derive the constraint and evolution equations which comprise the initial value formulation of Einstein's theory, and then in Sect. 16.3, we discuss the proof that this initial value problem is well-posed. Next, we examine methods for obtaining initial data sets which satisfy the Einstein constraint equations. In Sect. 16.4, we focus on the conformal method, in Sect. 16.5, we discuss the closely related conformal thin sandwich method, and in Sect. 16.6, we explore how such solutions can be obtained via gluing. Our discussion of the Einstein evolution equations is very brief; in Sect. 16.7, we comment primarily on globally hyperbolic solutions with singularities and the strong cosmic censorship conjecture, along with brief remarks concerning the stability of Minkowski spacetime, the Kerr solutions, and certain expanding cosmological solutions of the Einstein equations.

16.2 Derivation of the Einstein Constraint and Evolution Equations

To see how to build spacetime solutions of the Einstein field equations from initial data on a Riemannian manifold, it is useful to start with such a spacetime (M^4, g, Ψ), specify a spatial foliation of that spacetime (Such a foliation exists so long as the spacetime is globally hyperbolic; non globally hyperbolic spacetimes generally do not admit such foliations.), determine the geometric quantities which are well defined on the leaves of that foliation, and then calculate what Einstein equations (16.1) tell us about these geometric quantities. We do this here. (While earlier works of Lichnerowicz, of Dirac, and of Choquet-Bruhat explore the split of the spacetime metric into pieces defined relative to a spacelike foliation, and study the split of the Einstein equations into constraints and evolution equations, it is arguably the work of Arnowitt, Deser, and Misner (*ADM*) which is most responsible for introducing this approach to physicists. See, for example, [16.1].)

A three-dimensional hypersurface Σ^3, smoothly embedded in (M^4, g) (with embedding map $i : \Sigma^3 \to M^4$), is defined to be *spacelike* (or *spatial*) if the induced bilinear form $\gamma := i^* g$ on Σ^3 is Riemannian. Equivalently, the embedded hypersurface is spacelike if all vectors V tangent to $i(\Sigma^3)$ at p are spacelike with respect to g (so that $g(V, V) > 0$). A *spatial foliation* of the spacetime (M^4, g) is a smooth one-parameter family $i_t : \Sigma^3 \to M^4$ (for $t \in \mathbb{R}$) of embedded spacelike hypersurfaces such that each point $p \in M^4$ is contained in one and only one of the hypersurfaces $i_t(\Sigma^3)$ (referred to as the *leaves* of the foliation). We presume that it is also true that for each $p \in M^4$, there is one and only one point $q \in \Sigma^3$ such that $i_{\hat{t}}(q) = p$ for some $\hat{t}$.

We focus now on a particular leaf $i_t(\Sigma^3)$ of a chosen foliation i_t of a chosen spacetime (M^4, g). The definition of spatial foliation (above) guarantees that $i_t(\Sigma^3)$ is equipped with a Riemannian metric $\gamma_{[t]}$, which consequently defines a covariant derivative $\nabla_{[t]}$ and the corresponding Riemann, Ricci, and scalar curvature quantities, all intrinsic to $i_t(\Sigma^3)$. In addition to specifying these intrinsic geometric quantities, the foliation i_t equips $i_t(\Sigma^3)$ with extrinsic geometric quantities, including (i) $e_{\perp[t]}$, a future-pointing unit-length timelike vector field, orthogonal to all vectors tangent to $i_t(\Sigma^3)$; (ii) $\theta^{\perp}_{[t]}$, a unit-length one-form field such that (at each point in $i_t(\Sigma^3)$), $\langle \theta^{\perp}_{[t]}, e_{\perp[t]} \rangle = 1$ and $\langle \theta^{\perp}(p), V \rangle = 0$ for all vectors V tangent to $i_t(\Sigma^3)$; and $K_{[t]}$, the second fundamental form, defined by

$$K_{[t]}(U, V) := g(D_U e_{\perp[t]}, V) , \tag{16.2}$$

where D denotes the spacetime covariant derivative and where U and V are vector fields tangent to $i_t(\Sigma^3)$. It is straightforward to show that $K_{[t]}$ is a well-defined (spatial) tensor field with respect to the tangent spaces of $i_t(\Sigma^3)$, and further that it is symmetric. By contrast, $e_{\perp[t]}$ and $\theta^{\perp}_{[t]}$ are not spatially tensorial; rather, they act tensorially on the spacetime tangent spaces of M^4, restricted to the subset $i_t(\Sigma^3) \subset M^4$.

To relate the foliation-related quantities $\gamma_{[t]}$, $\nabla_{[t]}$, the spatial curvatures, $K_{[t]}, e_{\perp[t]}$ and $\theta^{\perp}_{[t]}$ to the spacetime geometry, it is useful to specify foliation-compatible bases. We obtain a (local in space) foliation-compatible coordinate basis by choosing coordinates $\{x_a\}|_{(a=1,2,3)}$ locally on Σ^3, using the foliation to transport these to its leaves $i_t(\Sigma^3)$, and adding the parameter t to produce the spacetime coordinates $\{x_a, t\} =: \{x_\alpha\}|_{(\alpha=1,2,3,0)}$. The corresponding dual coordinate bases $\{\partial_\alpha\}$ and $\{\mathrm{d}x^\alpha\}$ are foliation-compatible in the sense that $\mathrm{d}t$ annihilates vectors which are tangent to the leaves, ∂_t is transverse to the leaves, and ∂_a are tangent to the leaves; however, we note that ∂_t and $\mathrm{d}t$ need not be timelike everywhere (∂_t is timelike if $g(\partial_t, \partial_t) < 0$ and it is null if $g(\partial_t, \partial_t) < 0$; similarly $\mathrm{d}t$ is timelike or null or spacelike depending on the sign of $g^{-1}(\mathrm{d}t, \mathrm{d}t)$). Alternatively, one may choose the foliation-compatible basis $\{e_{\perp[t]}, \partial_a\}$ and its dual $\{\theta^{\perp}_{[t]}, \theta^a_{[t]} := \mathrm{d}x^a + M^a\,\mathrm{d}t\}$; here $M^a_{[t]}$ are the components of the *shift vector*, which is the spatial projection of ∂_t

$$\partial_t = N_{[t]} e_{\perp[t]} + M^a_{[t]} \partial_a . \tag{16.3}$$

The scalar $N_{[t]}$ is called the *lapse function*.

On the leaf $i_t(\Sigma^3)$, one easily verifies that the spacetime metric g is related to the spatial metric $\gamma_{[t]}$ by the identity (and its expansion)

$$g = \gamma_{[t]} - \theta^{\perp}_{[t]} \theta^{\perp}_{[t]} \tag{16.4}$$

$$= \gamma_{[t]ab} \left(\mathrm{d}x^a + M^a_{[t]}\,\mathrm{d}t\right)\left(\mathrm{d}x^b + M^b_{[t]}\,\mathrm{d}t\right) - N^2_{[t]}\,\mathrm{d}t^2 . \tag{16.5}$$

The expression relating the spacetime covariant derivative D (compatible with g) and the spatial covariant derivative $\nabla_{[t]}$ (compatible with $\gamma_{[t]}$) is not so easily derived; however, it does follow from the definition (16.2) of the second fundamental form that for any pair of vectors U and V tangent to $i_t(\Sigma^3)$, one has

$$D_U V = \nabla_{[t]U} V + K_{[t]}(U, V) e_{\perp[t]} , \tag{16.6}$$

thus identifying $K_{[t]}$ with the surface-normal projection of the spacetime covariant derivative. One also verifies from this definition, as well as from the Lie derivative identity $\mathcal{L}_Y g_{\alpha\beta} = D_\alpha Y_\beta + D_\beta Y_\alpha$, that

$$\mathcal{L}_{\partial_t}\gamma_{[t]ab} = 2N_{[t]}K_{[t]ab} + \mathcal{L}_{M_{[t]}}\gamma_{[t]ab}\,. \tag{16.7}$$

Notably, this result involves not only quantities on the slice $i_{[t]}(\Sigma^3)$, but also the rate of change of one of these quantities, $\gamma_{[t]}$, along the foliation.

It remains to consider the relationship between the spacetime curvature and the spatial curvatures. Since the calculations which lead to formulas for the components of the spacetime curvature tensor in terms of components of the spatial curvature tensor along with other quantities intrinsic to leaves of a chosen foliation are analogous to analyses first done (for Riemannian rather than for Lorentzian ambient geometries) by Gauss, Codazzi, and later Mainardi, these formulas are often referred to as the Gauss–Codazzi–Mainardi equations. Using

$$\mathcal{R}^\alpha{}_{\beta\gamma\delta} := \langle\theta^\alpha, D_{e_\gamma}D_{e_\delta}e_\beta - D_{e_\delta}D_{e_\gamma}e_\beta - D_{[e_\gamma,e_\delta]}e_\beta\rangle\,, \tag{16.8}$$

to represent the spacetime curvature tensor (with respect to spacetime dual bases $\{e_\alpha\}$ and $\{\theta^\beta\}$) and analogously

$$R^a{}_{bcd} := \langle\theta^a, \nabla_{e_c}\nabla_{e_d}e_b - \nabla_{e_d}\nabla_{e_c}e_b - \nabla_{[e_c,e_d]}e_b\rangle\,, \tag{16.9}$$

to represent the spatial curvature tensor (with respect to spatial dual bases $\{e_a\}$ and $\{\theta^b\}$), we find that these equations take the following form.

$$\mathcal{R}^a{}_{bcd} = R^a{}_{bcd} + K^a_c K_{bd} - K^a_d K_{bc}\,, \tag{16.10}$$
$$\mathcal{R}^a{}_{\perp cd} = \nabla_{e_c}K^a_d - \nabla_{e_d}K^a_c\,, \tag{16.11}$$

and

$$\mathcal{R}^a{}_{\perp b\perp} = -\mathcal{L}_{e_\perp}K^a_b + K^{am}K_{mb} + \frac{1}{N}\nabla^a\nabla_b N\,. \tag{16.12}$$

Note that, to avoid notational clutter, in equations (16.9)–(16.12) we have removed the "[t]" subscripts from foliation-defined quantities such as $\nabla_{[t]}$ and $K_{[t]ab}$ (The most straightforward method of deriving the Gauss–Codazzi–Mainardi equations is combining expression (16.6) relating D, ∇, and K with the expressions (16.8) and (16.9) for the curvatures. Details of these calculations appear in [16.2].).

If the spacetime (M^4, g, Ψ) under consideration satisfies the Einstein field equations (16.1) and if a spatial foliation $i_t: \Sigma^3 \to M^4$ has been specified, what do these equations tell us about the foliation-defined geometric quantities $\gamma_{[t]}$ and $K_{[t]}$? Combining the definition of the Einstein tensor $G_{\alpha\beta} = \mathcal{R}_{\alpha\beta} - \frac{1}{2}\mathcal{R}g_{\alpha\beta}$ with the Gauss–Codazzi–Mainardi equations (16.9)–(16.12), we readily obtain the following:

$$R + K_{cd}K^{cd} - (K^c_c)^2 = \tfrac{1}{2}G_{\perp\perp} = \frac{\kappa}{2}T_{\perp\perp}[\gamma, \psi, \pi]\,, \tag{16.13}$$
$$\nabla_c K^c_d - \nabla_d(K^c_c) = G_{\perp d} = \kappa T_{\perp d}[\gamma, \psi, \pi]\,, \tag{16.14}$$
$$\mathcal{L}_{e_\perp}K_{ab} + R_{ab} - 2K_{ac}K^c_b + K^c_c K_{ab} + \frac{1}{N}\nabla_a\nabla_b N = G_{ab} - \tfrac{1}{2}g_{ab}G^c_c = \kappa T_{ab}[\gamma, \psi, \pi] - \frac{\kappa}{2}g_{ab}T^c_c[\gamma, \psi, \pi]\,, \tag{16.15}$$

where (ψ, π) represent the nongravitational fields and their derivatives, after they have been decomposed into quantities which are well defined with respect to the leaves of the foliation. (For Maxwell's electromagnetic field, (ψ, π) correspond to (B_a, E_a).)

The first two sets of these equations, (16.13) and (16.14), have the very interesting feature that (presuming the quantity $T_{\perp\perp}[\gamma, \psi, \pi]$ behaves well), they involve quantities ($\gamma_{[t]}$ and $K_{[t]}$, not $e_\perp$) defined as tensors on the leaf $i_t(\Sigma^3)$, and do *not* directly involve any time derivatives of these quantities. By contrast, the remaining set of these equations, (16.15), does involve the time derivative of $K_{[t]}$.

There is nothing special about $i_t(\Sigma^3)$ or any other leaf of the foliation, nor is there anything special about the (arbitrarily) chosen foliation. Hence we see that on any spacelike hypersurface embedded in a spacetime (M^4, g, Ψ) which satisfies the Einstein equations, the surface quantities (γ, K, ψ, π) must satisfy (16.13) and (16.14). These are hence known as the *initial value constraint equations* for Einstein's theory.

The initial value problem (or Cauchy problem) for Einstein's theory turns things around. It addresses the following question: If we fix a three-dimensional manifold Σ^3 and if we choose a Riemannian metric γ_{ab} and a symmetric tensor K_{cd} (and possibly, nongravitational field data ψ and π on Σ^3 as well) which satisfy the constraint equations (16.13) and (16.14), is it always

true that there exists a spacetime (M^4, g, Ψ) which satisfies the (spacetime) Einstein field equations (16.1), and which contains an embedded submanifold $i(\Sigma^3) \subset M^4$ for which γ_{ab} is the induced metric i^*g_{ab} and K_{cd} is the induced second fundamental form? Further, if such a solution exists for a given set of initial data, is it unique in any sense? We address this question in the next section.

16.3 Well-Posedness of the Initial Value Problem for Einstein's Equations

It appears, from our discussion in Sect. 16.2, that the construction of a (vacuum) solution of the Einstein equations from initial data is a straightforward enterprise (especially with a computer readily at hand):

a) On a specified three-dimensional manifold Σ^3, we choose an initial data set (γ_{ab}, K_{cd}) which satisfies the (vacuum) constraint equations

$$R + K_{cd}K^{cd} - \left(K^c_c\right)^2 = 0\,, \qquad (16.16)$$

$$\nabla_c K^c_d - \nabla_d \left(K^c_c\right) = 0\,. \qquad (16.17)$$

b) We freely choose the lapse as a one-parameter (t) family of positive scalar functions $N(x, t)$ on Σ^3, and the shift as a one-parameter family of vector fields $M^a(x, t)$ on Σ^3.
c) Using the *evolution equations*

$$\mathcal{L}_{\partial_t}\gamma_{ab} = 2NK_{[t]ab} + \mathcal{L}_M\gamma_{ab}\,, \qquad (16.18)$$

$$\mathcal{L}_{\partial_t}K_{ab} = -NR_{ab} + 2NK_{ac}K^c_b - NK^c_cK_{ab} - \nabla_a\nabla_b N + \mathcal{L}_M K_{ab}\,, \qquad (16.19)$$

we evolve the initial data set into a one-parameter family of Riemannian metrics $\gamma_{ab}(x, t)$ and symmetric tensors $K_{ab}(x, t)$; and.
d) We construct the Lorentz metric

$$g = \gamma_{ab}(x, t)(\mathrm{d}x^a + M^a(x, t)\,\mathrm{d}t)(\mathrm{d}x^b + M^b(x, t)\,\mathrm{d}t) - N^2(x, t)\,\mathrm{d}t^2\,, \qquad (16.20)$$

on the spacetime manifold $M^4 = \Sigma^3 \times \mathbb{R}$ and verify that it is a solution of the vacuum Einstein equations $G_{\alpha\beta} = 0$.

To determine that this construction procedure in fact works, at least for some choices of the lapse and shift, one needs to prove a well-posedness theorem. Such a theorem was first proven for Einstein's theory during the 1950s by *Yvonne Choquet-Bruhat* [16.3]. This early work leaves the issue of uniquenness unsettled to a certain extent. However, the later work of *Choquet-Bruhat* with *Robert Geroch* [16.4] provides a strong form of uniqueness. Before stating this combined result, we find it useful to establish some terminology:

A spacetime (M^4, g) is *globally hyperbolic* if there exists an embedded spatial hypersurface $i(\Sigma^3) \subset M^4$ such that every future and past inextendible causal path intersects $i(\Sigma^3)$ once and only once (A smooth path $\eta : I \to M^4$ is *causal* if its tangent vector V is always either timelike $(g(V, V) < 0)$ or null $(g(V, V) = 0)$. Such a path is *inextendible* if there does not exist a smooth path $\tilde{\eta}$ which contains η as a proper subset.). If such a hypersurface exists, it is called a *Cauchy surface.* A spacetime is a *globally hyperbolic development* (*gh-development*) of a specified set of initial data (γ, K) on Σ^3 if

i) It as a solution of the Einstein field equations
ii) It is globally hyperbolic; and
iii) There exists an embedded hypersurface $\tilde{i}(\Sigma^3)$ such that (M^4, g) induces the specified initial data (γ, K) on $\tilde{i}(\Sigma^3)$ in the sense that $\gamma = \tilde{i}^*g$ and an appropriately modified version of (16.2) holds.

A spacetime (M^4, g) is a *maximal globally hyperbolic development* of (γ, K) on Σ^3 if every other gh-development of the same set of data is diffeomorphic to a subset of (M^4, g).

Using this terminology, we have the key result [16.3, 4]:

Theorem 16.1 Well-posedness of Einstein's vacuum equations
For any smooth set of initial data (γ, K) on Σ^3 which satisfies the vacuum constraint equations (16.16) and (16.17), there exists a unique (up to diffeomorphism) maximal globally hyperbolic development.

Before discussing how to prove this result and discussing ways in which the result can be generalized, we note a number of its features: First, Theorem 16.1 is fundamentally a local existence and uniqueness result. It guarantees that for any initial data set, there exist spacetimes which are gh-developments of that data, and

there exists a unique maximal gh-development, but it says nothing regarding whether that maximal development lasts for a finite or an infinite amount of proper (or coordinate) time. Second, Theorem 16.1 guarantees that there are globally hyperbolic spacetimes evolving from any given set of initial data, but it says nothing about nonglobally hyperbolic spacetimes containing an embedded hypersurface with that data. In many cases – the Taub-NUT spacetime [16.5] and the various generalized Taub-NUT spacetimes [16.6] are prominent examples – the maximal development of a given set of data can be extended smoothly across a Cauchy horizon to a nonglobally hyperbolic spacetime solution. A Cauchy horizon in a given spacetime (M^4, g) (*not* to be confused with a Cauchy surface, as discussed above) is a hypersurface $\mathcal{H}$ embedded in (M^4, g) which is null (i. e., at each point of $\mathcal{H}$, there is a vector tangent to $\mathcal{H}$ which is null) and which lies on the boundary between the region in (M^4, g) which is globally hyperbolic and the region which is not. The initial value problem tells us essentially nothing about whether or not such extensions exist for a given set of initial data. Third, Theorem 16.1 guarantees that there are choices of the lapse and shift which can be used in evolving initial data sets to produce gh-developments (numerical or otherwise), but it does not tell us whether this is true just for certain particular choices, or for all such choices. The proof of Theorem 16.1, as we see below, does provide some information regarding this issue, but does not resolve it. Finally, we note that while well-posedness theorems usually contain statements regarding continuity of the map from the domain space of initial data sets to the range space of solutions, this is not true of Theorem 16.1, as stated. Such a result does hold in an appropriate form for Einstein's equations; however, to avoid the unenlightening detail needed to state it, we do not include it here. We refer the interested reader to [16.7, Chap. 15].

The key for proving that the initial value problem for Einstein's field equations is well posed is to show that the system (16.1) can be, in a certain sense, transformed into a hyperbolic PDE system for which well-posedness is well established (A PDE system is *hyperbolic* if, roughly speaking, it has the characteristics of a system of wave equations. A precise definition of hyperbolic PDE systems is found, for example, in [16.8], and also in [16.7].). The vacuum Einstein equation system, which we can write as $\mathcal{R}_{\alpha\beta} = 0$, is not itself hyperbolic. The system

$$\mathcal{R}_{\alpha\beta} + \tfrac{1}{2}\mathcal{L}_Z g_{\alpha\beta} = 0\,, \tag{16.21}$$

with

$$Z_\alpha := -g^{\nu\mu}\left(\partial_\nu g_{\mu\alpha} - \tfrac{1}{2}\partial_\alpha g_{\nu\mu}\right) + \mathcal{F}_\alpha[g]\,, \tag{16.22}$$

for $\mathcal{F}_\alpha[g]$ a specified function of $g_{\alpha\beta}$ but *not* of its derivatives, takes the form

$$g^{\nu\mu}\partial_\nu\partial_\mu g_{\alpha\beta} = \mathcal{Q}_{\alpha\beta}[g, \partial g]\,, \tag{16.23}$$

and therefore *is* a hyperbolic PDE system for the metric $g_{\alpha\beta}$. This does not lead to the Einstein system itself being hyperbolic. However, using the Bianchi identities, one can show that the vector field Z^α satisfies the homogeneous hyperbolic system

$$g^{\nu\mu}\partial_\nu\partial_\mu Z^\alpha = -\mathcal{R}^\alpha_\beta Z^\beta\,, \tag{16.24}$$

so that the combined system of (16.21)–(16.24) is hyperbolic. Further, one can show (see, for example, [16.7, Chap. 14], or [16.9, Sect. 3.1]) that, so long as the geometric initial data (γ, K) satisfy the constraints, one can choose the corresponding initial values of $g_{\alpha\beta}$ and of $\partial_t g_{\alpha\beta}$ for the system (16.21)–(16.24) in such a way that the initial values of Z^α and $\partial_t Z^\alpha$ vanish. It then follows from standard results regarding (nonlinear) hyperbolic systems (see, for example, [16.7, Chap. 9]) that the system (16.21)–(16.24) is well posed; further, it follows that the solution $(g_{\alpha\beta}(x,t), Z^\alpha(x,t))$ has $Z^\alpha(x,t)$ vanishing for all time. Hence $(g_{\alpha\beta}(x,t), 0)$ is a solution of the vacuum Einstein equations, compatible with the initial data (γ, K).

We note that in the original proof [16.3] of the well posedness of the Einstein vacuum system, wave coordinates (then called *harmonic coordinates*) are used. Wave coordinates correspond to a particular choice of the function $\mathcal{F}_\alpha[g]$. Generalized wave coordinates, which play an important role in numerical simulations of black hole collisions [16.10] correspond to other choices of $\mathcal{F}_\alpha[g]$.

The techniques used to show that, for any given set of initial data (γ, K) satisfying the constraint equations, there exists a unique maximal spacetime development of that data, are very different from those (just discussed) which are used to prove local existence. To prove that a unique maximal development exists, the idea is to consider the set $\mathcal{M}_{[\gamma,K]}$ of all spacetime developments of (γ, K), and then define a partial ordering "$>$" on $\mathcal{M}_{[\gamma,K]}$, with $(\tilde{M}^4, \tilde{g}) > (M^4, g)$ if $(\tilde{M}^4, \tilde{g})$ is an extension of (M^4, g), up to diffeomorphism. One then shows that the ordered set $(\mathcal{M}_{[\gamma,K]}, >)$ has the properties needed to apply the maximality principle from set theory, from which it follows that $(\mathcal{M}_{[\gamma,K]}, >)$

has a unique maximal element (The maximality principle is equivalent to the axiom of choice. It is often referred to as *Zorn's lemma*, although some dispute the appropriateness of this label; see [16.7, Chap. 16]. One readily verifies that this maximal element is indeed a maximal spacetime development of the data (γ, K). Details of this argument can be found in [16.4] and in [16.7, Chap. 16].

There are two important ways in which one can generalize Theorem 16.1. The first of these involves coupling in nongravitational fields. It is *not* true that if a field theory has a well-posed initial value problem in Minkowski spacetime, then the standard (*comma-to-semicolon*) coupling of that field theory to Einstein's theory also necessarily has a well-posed initial value problem. (Roughly speaking, the *comma-to-semicolon* coupling of a given field theory to gravity involves replacing partial derivatives with metric-compatible covariant derivatives in the Lagrangian for the given field theory, multiplying this Lagrangian by the metric volume density $\sqrt{-\det g}$, adding the result to the Einstein Lagrangian $\mathcal{R}\sqrt{-\det g}$, and then varying the summed Lagrangian with respect to the metric and the fields.) The standard Klein–Gordon vector field theory, with field equations $D^\alpha D_\alpha W^\mu = m^2 W^\mu$, provides an example of a field theory which is well posed in flat spacetime, yet appears to be ill-posed if coupled to Einstein's theory. (It is generally difficult to *prove* that a system of partial differential equations does not have a well-posed Cauchy problem. However, for systems like the Klein–Gordon vector field theory with standard coupling to Einstein's theory, the presence of *derivative-coupling* terms in the equations causes standard analyses of hyperbolicity and well posedness to be essentially unmanageable; one is led to strongly suspect ill-posedness. Some of these issues are discussed in [16.11].) However, there are a number of field theories involving Einstein's theory coupled to nongravitational fields that do have well-posed Cauchy problems. These include Einstein–Maxwell, Einstein–Yang–Mills, Einstein–Dirac, certain forms of Einstein-scalar field theories, and certain forms of Einstein-fluid theories. (The verification of well posedness is straightforward for the Einstein–Dirac theory with commuting spinor fields. For Einstein–Dirac with anti-commuting spinor fields, and also for $N = 1$ supergravity (which requires that the spinor fields anti-commute), well-posedness holds, but in a subtle sense. See [16.12] for details.) For all of these, in addition to well posedness, one can prove that for each set of initial data satisfying the appropriate set of constraints, there is a unique maximal development spacetime solution of the coupled system.

The second way in which Theorem 16.1 can be generalized is by loosening the required degree of regularity. The original results, as well as the results we state here, presume that the initial data is smooth. Results of over forty years ago [16.13, 14] show that well posedness (as well as the existence of a unique maximal development) holds for initial data sets (γ, K) with γ contained in a local Sobolev space of index $s > 5/2$, and K in such a space with index $s - 1$. (The Sobolev index s indicates (weak) L^2 boundedness for order s derivatives of the indicated fields. See, e.g., [16.8] for definitions and properties of the Sobolev function spaces.) The work of *Klainerman* and *Rodnianski* [16.15] lowers the required regularity to $s > 2$, and the recent work of *Klainerman* et al. [16.16] indicates that (with certain restrictions) one can prove well posedness for $s = 2$ as well. We note that the drive to achieve lower regularity well-posedness results is motivated both by the desire to understand weak solutions, and by the use of low regularity results as tools in understanding long-time behavior of solutions.

Not stated as part of Theorem 16.1 above, but an important feature of Einstein's theory, is the fact that the quantities (γ_{ab}, K_{cd}) induced on any Cauchy surface in a spacetime solution of Einstein's equations must satisfy the constraint equations. Hence, as one evolves a spacetime solution from a set of initial data, the evolution equations (16.18) and (16.19) effectively preserve the constraints. Since numerical implementation of the Cauchy problem inevitably introduces small errors, the constraints are not precisely preserved in a numerically constructed solution. This appears to be a source of instability problems in numerical relativity.

16.4 The Conformal Method and Solutions of the Constraints

The crucial first step in building a spacetime solution of the Einstein equations via the initial value problem is to obtain a set of initial data which satisfies the constraints. Restricting our attention to the vacuum system for most of the discussion here, we presume that a fixed three-dimensional manifold Σ^3 has been chosen, and

we seek a Riemannian metric γ_{ab} on Σ^3 and a symmetric tensor K_{cd} on Σ^3 such that, together, (γ_{ab}, K_{cd}) satisfy the vacuum constraints (16.16) and (16.17).

There are three somewhat different goals one might have in studying solutions of the constraint equations:

i) The construction of physically interesting data sets, which may then be evolved into physically interesting spacetimes.
ii) The parametrization (in terms of an appropriate function space) of the set of all solutions of the constraint equations.
iii) The systematic study of various mathematical and physical properties of solutions of the constraints, including (for asymptotically Euclidean data sets) conserved quantities such as global mass and global angular momentum.

The most widely used analytical method for studying solutions of the constraints is the conformal method (together with the closely related conformal thin sandwich method). The conformal method is particularly adapted to the goal of finding a function space parametrization of the set of solutions of the constraints, the second goal listed above. It has also been used very effectively in the numerical construction of physically interesting initial data sets, especially those used to simulate black hole collisions (which are important for building gravitational signal templates). We discuss the conformal method here and the conformal thin sandwich method in Sect. 16.5; in Sect. 16.6, we discuss another approach for studying the constraints – gluing – which is especially effective as a tool for attaining the third goal listed above.

As a PDE system for the 12 functions encompassed in $\gamma_{ab}(x)$ and $K_{ab}(x)$, the four constraint equations (16.16) and (16.17) constitute an underdetermined system. The idea of the conformal system is to split the initial data into two sets – the *free (conformal) data*, and the *determined data* – in such a way that, for a specified choice of the free data, the constraint equations become a *determined* elliptical PDE system, to be solved for the determined data. There a number of ways to carry out this data split. We focus here on the *semi-decoupling split* (labeled in the early literature of the conformal method as *method A*; see [16.17]), which takes the following form:

- Free (*Conformal*) data:
  - λ_{ab} A Riemannian metric, specified up to conformal factor
  - σ_{ab} A divergence-free ($\nabla^a\sigma_{ab} = 0$), trace-free ($\lambda^{ab}\sigma_{ab} = 0$) symmetric tensor (the divergence and trace are defined here using the conformal metric λ_{ab})
  - τ A scalar field.
- Determined data:
  - ϕ A positive definite scalar field
  - W^a A vector field.

For a given choice of the free data, the four equations to be solved for the four functions of the determined data take the form

$$\nabla_m (LW)^m_a = \tfrac{2}{3}\phi^6 \nabla_a \tau\ , \tag{16.25}$$

$$\Delta\phi = \tfrac{1}{8}R\phi - \tfrac{1}{8}(\sigma^{mn} + LW^{mn})(\sigma_{mn} + LW_{mn})\phi^{-7} + \tfrac{1}{12}\tau^2\phi^5\ , \tag{16.26}$$

where the Laplacian Δ and the scalar curvature R are based on the λ_{ab}-compatible covariant derivative ∇_a, and where L is the corresponding conformal Killing operator, defined by

$$(LW)_{ab} := \nabla_a W_b + \nabla_b W_a - \tfrac{2}{3}\lambda_{ab}\nabla_m W^m\ . \tag{16.27}$$

Presuming that for the chosen conformal data one can indeed solve equations (16.25) and (16.26) for ϕ and W, then the initial data set (γ_{ab}, K_{cd}) constructed via the formulas

$$\gamma_{ab} = \phi^4 \lambda_{ab}\ , \tag{16.28}$$

$$K_{ab} = \phi^{-2}(\sigma_{ab} + LW_{ab}) + \tfrac{1}{3}\phi^4\lambda_{ab}\tau\ , \tag{16.29}$$

satisfies the Einstein constraint equations (16.16) and (16.17).

The four equations (16.25)–(16.26), which we collectively refer to as the LCBY equations (since their derivation is based on the work of *Lichnerowicz* [16.18], *Choquet-Bruhat*, and *York* [16.17]), are readily obtained by substituting formulas (16.28) and (16.29) into the vacuum constraints (16.16) and (16.17). Two key identities play a major role in the derivation of (16.25) and (16.26): The first is the formula for the scalar curvature of the conformally transformed metric $\gamma_{ab} = \phi^4\lambda_{ab}$, expressed in terms of the scalar curvature for λ_{ab} and derivatives of ϕ

$$R(\gamma) = \phi^{-4}R(\lambda) - 8\Delta_\lambda \phi\ . \tag{16.30}$$

We note that if we were to use a different power of ϕ as the conformal factor multiplying λ_{ab}, then this formula would involve squares of first derivatives of ϕ as

well. The second key formula relates the divergences of a traceless symmetric tensor ρ_{ab} as calculated using two different covariant derivatives – $\nabla_{(\gamma)}$ (compatible with the metric γ) and $\nabla_{(\lambda)}$ (compatible with λ)

$$\nabla^m_{(\gamma)}\rho_{mb} = \phi^{-2}\nabla^m_{(\lambda)}\left(\phi^2\rho_{mb}\right) . \qquad (16.31)$$

This formula dictates the choice of conformal scaling used in (16.29) for the trace-free part of K_{ab}; a different scaling would generally lead to the appearance of further $\nabla\phi$ terms in the LCBY equations, which are to be avoided.

Do the LCBY equations admit solutions (ϕ, W) for every choice of the conformal data $(\Sigma^3; \lambda_{ab}, \sigma_{cd}, \tau)$? It is easy to see that this is not the case: If we choose, for example, $\Sigma^3 = S^3$, λ_{ab} is the round metric on S^3, $\sigma_{cd} = 0$ everywhere, and $\tau = 1$ everywhere, then (16.25) takes the form $\nabla_m(LW)^m_a = 0$, which requires that LW_{ab} vanish everywhere. Equation (16.26) then takes the form $\Delta\phi = \frac{1}{8}R\phi + \frac{1}{12}\phi^5$. Since the right hand side of this equation is positive definite (recall the requirement that $\phi > 0$), it follows from the maximum principle on closed (compact without boundary) manifolds that there is no solution.

Since this example shows that solutions do not exist for every possible choice of the conformal data, we seek to determine exactly which sets of such data lead to a solution and which do not. This issue has been intensively studied for at least 40 years, and while much is known, there remains much to be determined. Roughly speaking, for conformal data with constant mean curvature (CMC), we generally know for which sets of conformal data $(\Sigma^3; \lambda_{ab}, \sigma_{cd}, \tau)$ solutions exist, and for which sets they do not. Uniqueness is well understood as well for CMC conformal data. For conformal data with nearly constant mean curvature, we know a number of classes of conformal data for which solutions exist, as well a number of classes for which they do not; there are, however, a number of unresolved cases. For conformal data with mean curvature far from constant, we know much less; however, there has been some progress recently in studying these sets of conformal data.

To explain more specifically what is known and what is not known about solving the LCBY equations, it is very useful to classify conformal data sets into a wide collection of classes, based on a number of criteria, which include the following:

- *Manifold type and asymptotic conditions*: Data on closed manifolds, asymptotically Euclidean data, asymptotically hyperbolic data, asymptotically conical data, asymptotically cylindrical data, or data on manifolds with boundary.
- *Regularity conditions*: Analytic data, smooth data, or data contained in specified Sobolev or Hölder spaces.
- *Coupled nongravitational fields*: Vacuum Einstein, Einstein–Maxwell, Einstein–Dirac, Einstein-scalar, Einstein-fluid, or Einstein–Vlasov.

It is beyond the scope of this chapter to discuss what is known for each of the several classes delineated by these criteria. Rather, we focus on what is known and what is not known for smooth data satisfying the vacuum Einstein constraints, either on closed manifolds or satisfying asymptotically Euclidean conditions. We comment briefly on some of other the other classes.

Within each class of conformal data, the key distinction is between those sets of data with constant mean curvature and those which have nonconstant mean curvature. The mean curvature of any initial data set is given by the function $\mathrm{tr}K := \gamma^{ab}K_{ab} = \tau$ (we use (16.28)–(16.29) to calculate the last equality here), so we see that the CMC condition corresponds to choosing conformal data with constant τ. Examining (16.25), we see that the CMC condition is important because constant τ implies the vanishing of the right hand side of equation (16.25), which then results in the vanishing of the tensor quantity $(LW)_{ab}$ (It follows from a straightforward analysis of the elliptic operator $\nabla \cdot L$ that in essentially all cases, if $\nabla_m(LW)^m_a = 0$, then LW_{ab} vanishes.). Consequently, for CMC conformal data the analysis of the LCBY equations reduces to that of the (decoupled) *Lichnerowicz equation*, which takes the form

$$\Delta\phi = \tfrac{1}{8}R\phi - \tfrac{1}{8}\sigma^{mn}\sigma_{mn}\phi^{-7} + \tfrac{1}{12}\tau^2\phi^5 . \qquad (16.32)$$

We note that for all of the nongravitational fields coupled to gravity which are listed above, the first of the LCBY equations takes the form $\nabla_m(LW)^m_a = \frac{2}{3}\phi^6\nabla_{\partial_a}\tau + J_a$, where J_a depends on the nongravitational conformal data, and does *not* involve ϕ (This effect occurs as a consequence of the choice one makes for the conformal rescaling of the nongravitational fields. Alternative choices could be made which would lead to J_a depending on ϕ, but there is little motivation for making such choices.). Consequently, while LW_{ab} does not vanish for CMC data if nongravitational fields are being considered, the LCBY equations *do* decouple, and as a result the determination if solutions

to the LCBY equations exist depends essentially on the Lichnerowicz equation (with extra terms corresponding to the nongravitational fields; e.g., in the case of the Einstein–Maxwell theory, the Lichnerowicz equation picks up a term of the form $-\rho\phi^{-3}$, where ρ is quadratic in the conformal electric and magnetic fields).

16.4.1 CMC Data on Closed Manifolds

As noted above, for a set of constant mean curvature conformal data $(\Sigma^3; \lambda_{ab}, \sigma_{cd}, \tau)$ for the vacuum Einstein equations, a solution to the LCBY equations exists (and a map to a solution of the constraint equations exists) if and only if there exists a solution to the Lichnerowicz equation (16.32). For this problem, we know exactly which sets of conformal data lead to solutions and which do not. To state these results (and to prove them), two features regarding conformal transformations of the conformal data are crucial.

First, we note the Yamabe theorem [16.19] for Riemannian metrics on closed manifolds, which states that every such metric can be conformally transformed to a metric of constant scalar curvature; further, for each metric the sign of that constant scalar curvature is unique. Hence, the set of Riemannian metrics on a given manifold Σ^3 are partitioned into three *Yamabe classes* $\mathcal{Y}^+$, $\mathcal{Y}^0$, and $\mathcal{Y}^-$, depending on that sign.

Second, we readily verify that the Lichnerowicz equation is *conformally covariant* in the following sense: If ϕ is a solution to the Lichnerowicz equation for a set of conformal data $(\Sigma^3; \lambda_{ab}, \sigma_{cd}, \tau)$, then for any smooth positive function θ, the function $\hat{\phi} = \theta^{-1}\phi$ is a solution to the Lichnerowicz equation for the conformal data $(\Sigma^3; \theta^4\lambda_{ab}, \theta^{-2}\sigma_{cd}, \tau)$. Combining this result with the Yamabe theorem, we see that *a solution to the Lichnerowicz equation exists for* $(\Sigma^3; \lambda_{ab}, \sigma_{cd}, \tau)$ *if and only if there exists a solution for the conformally transformed data* $(\Sigma^3; \theta^4\lambda_{ab}, \theta^{-2}\sigma_{cd}, \tau)$ *with* $R[\theta^4\lambda_{ab}] = +1, 0$ *or* -1, *depending upon the Yamabe class of* λ_{ab}.

Besides the Yamabe class of λ_{ab}, two aspects of a given set of conformal data determine whether or not a solution to the Lichnerowicz equation exists. One of them is whether or not the constant τ is zero or not. In table of results (Table 16.1), we label these alternatives $\tau = 0$ or $\tau \neq 0$. The other is whether or not the function $\sigma_{cd}\sigma^{cd}$ is identically zero (on Σ^3) or not. We label these alternatives $\sigma \equiv 0$ or $\sigma \not\equiv 0$. Using Y to indicate that solutions exist for conformal data sets in a certain class and using N to indicate that they do not, we summarize the results in Table 16.1.

Table 16.1 Solvability of Lichnerowicz equation for CMC data

	$\sigma \equiv 0$, $\tau = 0$	$\sigma \not\equiv 0$, $\tau = 0$	$\sigma \equiv 0$, $\tau \neq 0$	$\sigma \not\equiv 0$, $\tau \neq 0$
$\mathcal{Y}^+$	N	Y	N	Y
$\mathcal{Y}^0$	Y	N	N	Y
$\mathcal{Y}^-$	N	N	Y	Y

We note that for all those data sets such that solutions exist, except for that class of data with $\lambda \in \mathcal{Y}^0$, with $\sigma \equiv 0$ and with $\tau = 0$, the solutions are unique. In the latter (somewhat trivial) case, any (positive) constant ϕ is a solution.

The results summarized in the table here were to a large extent proven in the 1970s by *Yvonne Choquet-Bruhat*, *James York*, and *Niall O'Murchadha*; see [16.17] for a discussion of this early work. The complete proof, including two cases not handled by the previous work, appears in [16.20]. There it is shown that the *No* cases can all be proven using the *maximum principle*, stated in the following form: On a closed manifold, the equation $\Delta\phi = f(x, \phi)$, with $f(x, \phi)$ either nonvanishing and nonpositive or nonvanishing and nonnegative, has no solution. It is also shown there that the *Yes* cases can all be proven using the *sub and super solution theorem*, which can be stated as follows: If there exist a pair of positive functions $\phi_+ \geq \phi_-$ which satisfy the inequalities

$$\Delta\phi_- \leq \tfrac{1}{8}R\phi_- - \tfrac{1}{8}\sigma^{mn}\sigma_{mn}\phi_-^{-7} + \tfrac{1}{12}\tau^2\phi_-^5\,, \quad (16.33)$$

and

$$\Delta\phi_+ \geq \tfrac{1}{8}R\phi_+ - \tfrac{1}{8}\sigma^{mn}\sigma_{mn}\phi_+^{-7} + \tfrac{1}{12}\tau^2\phi_+^5\,, \quad (16.34)$$

then there exists a solution ϕ of the Lichnerowicz equation (16.32), with $\phi_+ \geq \phi \geq \phi_-$. For some of the six *Yes* cases (specifically those with the metric in the negative Yamabe class), constant sub and super solutions are easily found. For others, nonconstant sub and super solutions are needed, and are not so easily found.

We can use the results summarized in the table, along with the conformal covariance stated above and a certain scaling invariance, to *parametrize* the set of CMC solutions of the vacuum constraints on a chosen closed manifold Σ^3. We first note that every solution (γ_{ab}, K_{ab}) of the constraint equations (16.16) and (16.17) can be obtained using the conformal method, since one can choose $\lambda_{ab} = \gamma_{ab}$, $\tau = K^c_{\ c}$, and σ_{cd} equal to the divergence-free trace-free projection (see [16.17]) of K_{cd}; with this conformal data, $W^a = 0$ and $\phi = 1$ satisfy the LCBY equations, leading us

back to (γ_{ab}, K_{ab}). Next, we note that the conformal covariance result implies that for any positive function θ the two sets of conformal data $(\Sigma^3; \lambda_{ab}, \sigma_{cd}, \tau)$ and $(\Sigma^3; \theta^4\lambda_{ab}, \theta^{-2}\sigma_{cd}, \tau)$ lead to related solutions of the Lichnerowicz equation and thence to *identical* solutions (γ_{ab}, K_{cd}) of the constraint equations (16.16) and (16.17). The scaling result [16.21] states that for a positive constant A, the two sets of CMC conformal data $(\Sigma^3; \lambda_{ab}, \sigma_{cd}, \tau)$ and $(\Sigma^3; A^2\lambda_{ab}, A\sigma_{cd}, A^{-1}\tau)$ admit the same solution ϕ of the Lichnerowicz equation. These lead to different solutions (γ_{ab}, K_{cd}) and $(A^2\gamma_{ab}, A\ K_{cd})$ of the constraint equations, but these solutions are related by an essentially trivial scaling. As a consequence of these considerations, we determine that for any chosen closed manifold Σ^3, the set of all conformal data sets marked *Yes* in the table, quotiented out by the conformal covariance equivalence and by the scaling equivalence, provides a faithful (bijective) parametrization of the space of CMC solutions of the vacuum constraints on Σ^3.

These results for CMC solutions of the vacuum constraints are essentially replicated if one considers the Einstein–Maxwell, Einstein–Yang–Mills, Einstein–Dirac, or Einstein-fluid constraint equations [16.22] (For the Einstein-fluid constraints, there are consistent choices of conformal scaling of the fluid fields which lead to a more difficult set of LCBY equations, analogous to those arising for certain Einstein-scalar theories. Conformal scalings can, however, always be chosen in a way which avoids these problems [16.22].). On the other hand, the results for the Einstein-scalar field theories with certain types of field theory potentials are not nearly as complete. Details of the difficulties that can arise are discussed in [16.23]; see also [16.24], which explores how to handle some of these difficulties.

16.4.2 Asymptotically Euclidean CMC Data

The decoupling of the LCBY equations which results from the assumption of constant mean curvature does not depend upon the topology of Σ^3 or the asymptotic conditions of the conformal data. Hence, as for the closed manifold case discussed above, the analysis of the conformal method for asymptotically Euclidean data (as well as for asymptotically hyperbolic data or any other chosen asymptotic condition) reduces to the study of the solvability of the Lichnerowicz equation. We note that the fall-off conditions which define a set of data to be asymptotically Euclidean (see, for example, [16.25] or [16.26] for a precise statement of these fall-off conditions) require that the mean curvature, if constant, be zero. Hence asymptotically Euclidean data which is CMC must be maximal; i. e., $K^c_c = \tau = 0$.

The criterion for the Lichnerowicz equation to admit a solution for asymptotically Euclidean conformal data is independent of σ_{cd}, and depends only the metric. As proven by *Cantor* [16.27], the Lichnerowicz equation admits a solution for a given set of asymptotically Euclidean conformal data if and only if the metric admits a conformal transformation which results in the scalar curvature vanishing everywhere. Such metrics are called *Yamabe positive*, and *Brill* and *Cantor* [16.28] have also shown (with a correction pointed out by *Maxwell* [16.29]; see also [16.30]) that an asymptotically Euclidean metric is Yamabe positive in the sense just described if and only if for every nonvanishing compactly supported function f on Σ^3, the inequality

$$\inf_{\{f \neq 0\}} \frac{\int_M (|\nabla f|^2 + Rf^2)\sqrt{\det \lambda}}{||f||^2_{L^2}} > 0 \tag{16.35}$$

holds.

We note that the same criterion for the solvability of the Lichnerowicz equation holds for asymptotically Euclidean data for the Einstein–Maxwell, Einstein–Yang–Mills, Einstein–Dirac, Einstein-fluid, and other such theories involving nongravitational fields coupled to Einstein's theory. We also note, without elaboration, results concerning the solvability of the Lichnerowicz equation for asymptotically hyperbolic data in [16.31], and results for asymptotically cylindrical data sets in [16.32] and [16.33].

16.4.3 Near-CMC Data

If we drop the non-CMC condition on the choice of the conformal data, then $\nabla\tau$ does not vanish, and we must deal with the fully coupled LCBY system. While not much is known about the solvability of the LCBY system for general sets of non-CMC conformal data, for those sets of data with $|\nabla\tau|$ small in some appropriate sense, one can in many cases determine whether or not solutions exist.

One way to analyze the coupled system is to use the iterated Gummel method, which replaces (16.25) and (16.26) by the sequence of semidecoupled PDE systems

$$\nabla_c (LW_{(n)})^c_a = \tfrac{2}{3}\phi^6_{(n-1)}\nabla_a\tau \tag{16.36}$$

$$\begin{aligned}\Delta\phi_{(n)} = {} & \tfrac{1}{8}R\phi_{(n)} - \tfrac{1}{8}\left(\sigma^{cd} + LW^{cd}_{(n)}\right) \\ & \times (\sigma_{cd} + LW_{(n)cd})\phi^{-7}_{(n)} + \tfrac{1}{12}\tau^2\phi^5_{(n)}\ .\end{aligned} \tag{16.37}$$

The idea is to:

i) Choose an initializing value for ϕ_0;
ii) Show that a sequence $(\phi_{(n)}, W_{(n)})$ of solutions to (16.36)–(16.37) exists;
iii) Prove that there are uniform upper and lower bounds for the sequence $(\phi_{(n)}, W_{(n)})$; and
iv) Use those upper and lower bounds together with a contraction mapping argument to show that the sequence $(\phi_{(n)}, W_{(n)})$ converges uniformly to a limit (ϕ, W), which solves the LCBY equations.

The assumption that $|\nabla\tau|$ is small plays a crucial role in carrying out steps (iii) and (iv). It does this because, as a consequence of the elliptic equation (16.25) and its sequential analog (16.36), the norm of LW is controlled by $|\nabla\tau|\phi^6$. Thence the term involving the square of LW in (16.26), which contains factors of $(\phi^6)^2$ times ϕ^{-7} – i. e., ϕ^5 – competes with the $\tau^2\phi^5$ term. To ensure that $\tau^2\phi^5$ (which has the favorable sign) wins this competition and to thereby retain the control of ϕ that one has in the CMC case, it is sufficient to require that $\frac{|\nabla\tau|}{|\tau|}$ be sufficiently small.

Analyses of the sort discussed above have been carried out in [16.34, 35]. Combining these results with others proven in [16.36] and [16.37], one has the following table which lists, for each of 12 classes of near-CMC conformal data on closed manifolds, whether solutions to the LCBY equations are known to exist (Y), are known to not exist (N), or are not known either way ($?$) (In all of the Y cases, uniqueness holds as well.). In Table 16.2, the Yamabe classes and their labels $\{\mathcal{Y}^+, \mathcal{Y}^0, \mathcal{Y}^-\}$ are the same as used above for the CMC table, and the labels $\sigma \equiv 0$ and $\sigma \not\equiv 0$ denoting whether or not $|\sigma|$ is identically zero or not are also the same as for the CMC table. For the function τ, the notation $\tau^2 > 0$ indicates those sets of conformal data with τ (presumed smooth) having no zeros, while $\tau^2 \not> 0$ labels those sets of conformal data with τ allowed to have zeroes.

Not surprisingly, the classes which cause the most difficulty are those in which τ has zeroes; we have a fairly complete understanding of whether or not solutions exist for the classes in which τ is bounded away from zero.

Near-CMC conformal data sets which are asymptotically Euclidean have also been studied. Techniques similar to those discussed briefly here show [16.26] that the criterion for the existence of solutions to the LCBY equations for near-CMC asymptotically Euclidean conformal data sets is very similar to that for CMC data. Roughly the same holds true for asymptotically hyperbolic data sets; see [16.38].

Table 16.2 Solvability of conformal constraints for near-CMC data

	$\sigma \equiv 0$, $\tau^2 \not> 0$	$\sigma \not\equiv 0$, $\tau^2 \not> 0$	$\sigma \equiv 0$, $\tau^2 > 0$	$\sigma \not\equiv 0$, $\tau^2 > 0$
$\mathcal{Y}^+$	?	Y	N	Y
$\mathcal{Y}^0$	?	Y	N	Y
$\mathcal{Y}^-$	?	?	Y	Y

16.4.4 Far-CMC Data

Without either the CMC or the near-CMC condition imposed, the analysis of the LCBY equations is considerably more difficult. For most sets of conformal data satisfying neither of these conditions (we use the label *far-CMC* both for initial data sets $(\Sigma^3; \gamma_{ab}, K_{cd})$ and for conformal data sets $(\Sigma^3; \lambda_{ab}, \sigma_{cd}, \tau)$ which are not CMC and satisfy no smallness condition on K^c_c or on τ), it is not known whether or not solutions exist.

There has, however, been recent work which begins to explore whether or not solutions exist for far-CMC data, and also explores their multiplicity. We briefly discuss three of these recent works here.

The work of *Holst* et al. [16.39] together with the important follow-up work of *Maxwell* [16.40] effectively swaps the near-CMC condition of smallness of $|\nabla\tau|$ for a condition requiring the smallness of $|\sigma_{cd}|$. More specifically, their combined work shows that the LCBY equations admit a solution for a given set of conformal data on a closed manifold if the following conditions hold:

i) $\lambda_{ab} \in \mathcal{Y}^+$
ii) λ_{ab} does not admit a conformal Killing field
iii) $|\sigma_{cd}|$ is sufficiently small
iv) σ_{cd} is not identically zero.

We emphasize the fact that these conditions include *no* restriction on the mean curvature function τ. We also note that while these conditions guarantee existence, they tell us nothing about uniqueness.

Besides showing that a small but interesting class of far-CMC conformal data is mapped to solutions of the constraint equations, this work introduces a new analytical tool to the study of the conformal method. The proof that the LCBY equations admit solutions for sets

of conformal data of this sort described above relies on expressing solutions of the LCBY equation formally as fixed points of a map

$$\mathcal{F}:\phi\to\Delta^{-1}\times\left[\tfrac{1}{8}R\phi-\tfrac{1}{8}\left(\sigma+L\{(\nabla\cdot L)^{-1}\left(\tfrac{2}{3}\phi^{6}\nabla\tau\right)\}\right)^{2}\phi^{-7}+\tfrac{1}{12}\tau^{2}\right], \tag{16.38}$$

and proving that such fixed points exist using Schauder compactness.

Schauder techniques also play a role in the work *Dahl* et al. [16.37]. They show that for certain classes of conformal data – data on a closed manifold Σ^3 with the metric (any Yamabe class) admitting no conformal Killing fields, with τ bounded away from zero, and with $|\sigma|^2$ not identically zero if $\lambda_{ab}\in\mathcal{Y}^-$ – solutions of the LCBY equations exist if the equation

$$\Delta Y^a=|LY|\frac{1}{\tau}\nabla^a\tau\,, \tag{16.39}$$

does *not* admit a solution. Thus, from this perspective, one proves the existence of LCBY solutions for a given set of conformal data by showing that (16.39) (called the *limit equation* by the authors of [16.37]) admits no solution. This approach does not directly prove the existence of solutions of the LCBY equation for sets of far-CMC conformal data, but it can in principle be used to do this. In [16.41], it shown that an analogous limit equation-type result holds for asymptotically hyperbolic conformal data sets.

It is interesting that the derivation of these limit equation results relies on the study of solutions of a family of *deformed LCBY*$_\epsilon$ *equations* in which (16.26) is left unchanged, but (16.25) is replaced by

$$\nabla_m(LW)^m_a=\tfrac{2}{3}\phi^{(6-\epsilon)}\nabla_a\tau\,. \tag{16.40}$$

It is much easier to prove that solutions to the system (16.26)–(16.40), with $\epsilon>0$, exist, than to prove the existence of solutions to the LCBY equations. The idea then is to study the $\epsilon\to0$ limit of solutions of the LCBY$_\epsilon$ system, and use the analysis of these limits to shed light on the LCBY system. Such an analysis leads to the limit equation results.

The third work concerning solutions of the LCBY equations for far-CMC conformal data which we discuss here is that of *Maxwell* [16.42], in which he studies a very simple class of planar symmetric conformal data and finds somewhat surprising results. The conformal data sets he considers are characterized as follows:

i) $\Sigma^3=T^3$.
ii) The metric is flat.
iii) σ_{cd} has planar (T^2) symmetry, and therefore, as a consequence of the trace-free and divergence-free conditions, is a matrix with constant entries (two free constants).
iv) τ is a step function, with a pair of discontinuities.

Working with these simple data sets, which are parametrized by a small set of constants, Maxwell can directly (numerically) construct solutions, if they exist. He finds that for certain ranges of values of the parameters, no solutions exist; for other ranges, multiple solutions exist; and finally for a third range, unique solutions can be obtained.

It is very intriguing to consider whether these results of Maxwell generalize to conformal data sets without any discontinuities, and to data sets with less imposed symmetry. In any case, these results suggest that the behavior of the conformal method for far-CMC conformal data sets could be interesting and complicated.

16.5 The Conformal Thin Sandwich Method

The conformal method has proven to be a remarkably useful tool for generating and parametrizing and analyzing solutions of the Einstein constraint equations. It does, however, have some minor drawbacks:

a) The conformal data is somewhat remote from the physical data, since the conformal factor changes the physical scale on different regions of space.
b) While casting the constraints into a determined PDE form has the advantage of producing PDEs of a relatively familiar (elliptic) form, one does give up certain flexibilities which are inherent in an underdetermined set of PDEs.
c) In choosing a set of conformal data, one has to first project out a divergence-free trace-free tensor field (σ_{cd}).

d) While the LCBY system is conformally covariant in the sense discussed above in Sect. 16.4.1 for CMC conformal data, this is not the case for non-CMC conformal data.

The last two of these problems can be removed by modifying the conformal method in a way which *York* [16.43] has called the *conformal thin sandwich* (CTS) approach. The basic idea of the conformal thin sandwich approach is essentially the same as that of the conformal method. There are, however, two important differences. First, the CTS free data sets are larger than the free (conformal) data sets of the conformal method in the following sense: Like the conformal data sets, the CTS data sets include a conformal metric λ_{ab} and a mean curvature scalar τ. In addition, the CTS data sets include a trace-free tensor U_{cd} to replace the divergence-free trace-free tensor σ_{cd} of the conformal data sets, plus an extra scalar field η. Second, after solving the following set of CTS equations (analogous to the LCBY equations)

$$\begin{aligned}&\nabla_m((2\eta)^{-1}(LX))_a^m\\&=\tfrac{2}{3}\Phi^6\nabla_a\tau+\nabla_m((2\eta)^{-1}U_a^m)\,,\end{aligned}\tag{16.41}$$

$$\begin{aligned}\Delta\Phi&=\tfrac{1}{8}R\Phi\\&-\tfrac{1}{8}(U^{mn}+LY^{mn})\\&\times(U_{mn}+LY_{mn})\Phi^{-7}+\tfrac{1}{12}\tau^2\Phi^5\,,\end{aligned}\tag{16.42}$$

for the conformal factor Φ and the vector field Y^a, one constructs not just the initial data (γ_{ab}, K_{cd}), but the lapse N and the shift M^a as well,

$$\gamma_{ab}=\Phi^4\lambda_{ab}\tag{16.43}$$

$$K_{ab}=\Phi^{-2}(-U_{ab}+LY_{ab})+\tfrac{1}{3}\Phi^4\lambda_{ab}\tau\tag{16.44}$$

$$N=\Phi^6\eta\tag{16.45}$$

$$M^a=Y^a\,.\tag{16.46}$$

It is clear that in using the CTS approach, one need not project out a divergence-free part of a symmetric trace-free tensor. As well also, one also readily checks that the CTS method is conformally covariant in the sense discussed above: the initial data (γ_{ab}, K_{cd}) and the lapse and shift (N, M^a) generated from the CTS data set $(\lambda_{ab}, U_{ab}, \tau, \eta)$ and from the CTS data set $(\theta^4\lambda_{ab}, \theta^{-2}U_{ab}, \tau, \theta^6\eta)$ are identical. Furthermore, since the mathematical form of (16.41) and (16.42) is very similar to that of (16.25) and (16.26), the solvability results for the conformal method can be essentially carried over to the CTS approach.

There is, however, one problematic feature of the conformal thin sandwich approach. The problem arises if we seek CMC initial data with the lapse function chosen so that the evolving data continues to have constant mean curvature. (Such a gauge choice is often used in numerical relativity.) In the case of the conformal method, after solving (16.25) and (16.26) to obtain initial data (γ_{ab}, K_{cd}) which satisfies the constraints, one achieves this by proceeding to solve a linear homogeneous elliptic PDE for the lapse function. One easily verifies that solutions to this extra equation always exist. By contrast, in the CTS approach, the extra equation takes the form

$$\begin{aligned}\Delta(\Phi^7\eta)&=\tfrac{1}{8}\Phi^7\eta R+\tfrac{5}{2}(\Phi\eta)^{-1}(U-LY)^2\\&+\Phi^5Y^m\nabla_m\tau-\Phi^5\,,\end{aligned}\tag{16.47}$$

which is coupled to the system (16.41) and (16.42). The coupling is fairly intricate; hence little is known about the existence of solutions to the system, and it has been seen that there are problems with uniqueness. These difficulties do not arise, of course, if one makes no attempt to preserve the constant mean curvature condition.

16.6 Gluing Solutions of the Constraint Equations

Both the conformal method and the conformal thin sandwich method are procedures for generating initial data sets which satisfy the Einstein constraint equations from scratch. The gluing procedures, which we discuss here, produce new solutions of the constraint equations by combining existing solutions. While the gluing procedures have not yet turned out to be as useful as the conformal method and CTS method for the practical generation of physical interesting initial data sets, they have proven to be very effective for certain applications and for settling certain conjectures. We outline some of these applications below, after describing the two gluing procedures which have been developed, and how they work.

The *asymptotic exterior gluing*, developed by *Corvino* and *Schoen* [16.44, 45], works as follows. We presume that $(\Sigma^3, \gamma_{ab}, K_{cd})$ is an asymptotically Euclidean initial data set which satisfies the Einstein

constraints, and also satisfies certain asymptotic conditions (as specified in [16.45]). For any compact region $\Xi^3 \subset \Sigma^3$ for which $\Sigma^3 \setminus \Xi^3 = \mathbb{R}^3 \setminus B^3$ (where B^3 is a ball in $\mathbb{R}^3$), there is a smooth asymptotically Euclidean solution of the constraints on Σ^3 which is identical to the original solution on $\Xi^3 \subset \Sigma^3$, and is identical to Cauchy data for the Kerr solution on $\Sigma^3 \setminus \tilde{\Xi}^3$ for some $\tilde{\Xi}^3 \subset \Sigma^3$. In words, this technique allows one to smoothly glue any interior region of an asymptotically Euclidean solution to an exterior region of a slice of a Kerr solution. There is generally a transition zone between the interior chosen region and the exterior Kerr region $\Sigma^3 \setminus \Xi$ which is unknown, but is a solution of the constraint equations. We note that for asymptotically Euclidean solutions of the constraints with $K_{cd} = 0$, this method glues any interior region to an exterior region of a slice of the Schwarzschild solution.

In some sense, the Corvino–Schoen asymptotic exterior gluing result is very surprising. *If* the constraint equations were a determined elliptic system, one would *not* expect to be able to smoothly glue two solutions together like this, even with a transition region (satisfying the constraints). The key to proving that asymptotic exterior gluing indeed works is the exploitation of the underdetermined character of the constraints as a PDE system. Some of the ideas developed in the Corvino–Schoen work have also proven to be useful for localizing metric deformations as solutions of the constraints, as shown in the work of *Chruściel* and *Delay* [16.46].

The other gluing procedure, *connected sum gluing*, was developed first by *Isenberg* et al. [16.47] with further work done together with *Chruściel* et al. [16.48] and with *Maxwell* et al. [16.22]. The idea here is to start with a pair of solutions of the (vacuum) constraints $(\Sigma_1^3, \gamma_1, K_1)$ and $(\Sigma_2^3, \gamma_2, K_2)$ and to choose of a pair of points $p_1 \in \Sigma_1^3$ and $p_2 \in \Sigma_2^3$, one point contained in each solution. Based on these solutions, connected-sum gluing produces a new set of initial data $(\Sigma^3_{(1-2)}, \gamma_{(1-2)}, K_{(1-2)})$ with the following properties:

i) $\Sigma_{(1-2)}$ is diffeomorphic to the connected sum $\Sigma_1^3 \# \Sigma_2^3$ (which is constructed by first removing a ball from each of the manifolds Σ_1^3 and Σ_2^3, and then using a cylindrical bridge $S^2 \times I$ (where I is an interval in R^1) to connect the resulting S^2 boundaries on each manifold).
ii) $(\Sigma^3_{(1-2)}, \gamma_{(1-2)}, K_{(1-2)})$ is a solution of the constraints everywhere on $\Sigma^3_{(1-2)}$.
iii) On that portion of $\Sigma^3_{(1-2)}$ which corresponds to $\Sigma_1^3 \setminus \{\text{ball around } p_1\}$, the data $(\gamma_{(1-2)}, K_{(1-2)})$ is isomorphic to (γ_1, K_1), with a corresponding property holding on that portion of Σ_2^3 which corresponds to $\Sigma_2^3 \setminus \{\text{ball around } p_2\}$.

Connected sum gluing can be carried out for fairly general sets of initial data. The sets may be asymptotically Euclidean, asymptotically hyperbolic, specified on a closed manifold, or indeed anything else. The only condition that the data sets must satisfy is that, in sufficiently small neighborhoods of each of the points at which the gluing is to be done, there do not exist nontrivial solutions ξ to the equation $D\Theta^*_{(\gamma,K)}\zeta = 0$, where $D\Theta^*_{(\gamma,K)}$ is the operator obtained by taking the adjoint of the linearized constraint operator. (If a solution to this equation does exist on some region $\Lambda \in \Sigma^3$, it follows from the work of Moncrief that the spacetime development of the data on Λ admits a nontrivial isometry.) In work by *Beig* et al. [16.49], it is shown that this condition (sometimes referred to as *No KIDs*, meaning *no (localized) Killing initial data*) is indeed generically satisfied.

The details of the proof that connected sum gluing can be carried out as generally as described above are beyond the scope of this chapter; see [16.48] along with the references cited in that work for a complete discussion. We do wish to note three features of the proof: First, the proof is constructive in the sense that it outlines a systematic, step-by-step mathematical procedure for doing the gluing. In principle, one should be able to carry out the gluing procedure numerically. Second, connected sum gluing relies primarily on the conformal method, but it also requires the use of a nonconformal deformation (dependent on the techniques of Corvino and Schoen, and of Chruściel and Delay), so as to guarantee that the glued data is not just very close to the given data on regions away from the connecting bridge, but is indeed identical to it. Third, while the Corvino–Schoen asymptotic exterior gluing has not yet been proven to work for solutions of the constraints with source fields, connected sum gluing (up to the last step, which relies on Corvino–Schoen) has been shown to work for most matter source fields of interest [16.22]. It has also been shown to work for general dimensions greater than or equal to 3.

As noted above, while gluing has not seen widespread use as a procedure for producing physically interesting initial data sets, it has proven to be very valuable for a number of applications. We note a collection of these applications here:

1. *Spacetimes with regular asymptotic structure*: Until recently, it was not known whether there is

a large class of spacetime solutions of the Einstein equations which admit the conformal compactification and consequent asymptotically simple structure at null and spacelike infinity characteristic of the Minkowski and Schwarzschild spacetimes. Using asymptotic exterior gluing, together with Friedrich's analyses of spacetime asymptotic structures and arguments of *Chruściel* and *Delay* [16.50], one can produce such a class of solutions.

2. *Initial data for the gravitational N-body problem*: To model the physics of a system consisting of N chosen astrophysical bodies interacting gravitationally, it is important to be able to construct initial data sets which solve the Einstein constraints and which accurately model the bodies of interest, their initial placement, and their initial momenta, all in a single asymptotically Euclidean space. *Chruściel* et al. [16.51] have used gluing techniques to show that for any chosen set of N asymptotically Euclidean solutions of the constraints representing black holes, stars, or other astrophysical objects of interest, one can construct a new asymptotically Euclidean solution which includes interior regions of these N chosen solutions, placed as desired (so long as the distances between the bodies are sufficiently large) and with the desired relative momenta.
3. *Adding a black hole to a cosmological spacetime*: Although there is no clear established definition for a black hole in a spatially compact solution of Einstein's equations, one can glue an asymptotically Euclidean solution of the constraints to a solution on a compact manifold, in such a way that there is an apparent horizon on the connecting bridge. Studying the nature of these solutions of the constraints, and their evolution, could be useful in trying to understand what one might mean by a black hole in a cosmological spacetime.
4. *Adding a wormhole to your spacetime*: While we have discussed connected sum gluing as a procedure which builds solutions of the constraints with a bridge connecting two points on different manifolds, it can also be used to build a solution with a bridge connecting a pair of points on the *same* manifold. This allows one to do the following: If one has a globally hyperbolic spacetime solution of Einstein's equations, one can choose a Cauchy surface for that solution, choose a pair of points on that Cauchy surface, and glue the solution to itself via a bridge from one of these points to the other. If one now evolves this glued-together initial data into a spacetime, it will likely become singular very quickly because of the collapse of the bridge. Until the singularity develops, however, the solution is essentially as it was before the gluing, with the addition of an effective wormhole. Hence, this procedure can be used to glue a wormhole onto a generic spacetime solution.
5. *Removing topological obstructions for constraint solutions*: We know that every closed three dimensional manifold Σ^3 admits a solution of the vacuum constraint equations. To show this, we use the fact that Σ^3 always admits a metric Γ of constant negative scalar curvature. One easily verifies that the data $(\gamma = \Gamma, K = \Gamma)$ is a CMC solution. Combining this result with connected sum gluing, one can show that for every closed Σ^3, the manifold $\Sigma^3 \setminus \{p\}$ admits both an asymptotically Euclidean and an asymptotically hyperbolic solution of the vacuum constraint equations.
6. *Proving the existence of vacuum solutions on closed manifolds with no CMC Cauchy surface*: Based on the work of *Bartnik* [16.52] one can show that if one has a set of initial data on the manifold $T^3 \# T^3$ with the metric components even-reflective across a central sphere and the components of K odd-reflective across that same central sphere, then the spacetime development of that data does not admit a CMC Cauchy surface. Using connected sum gluing, one can show that indeed initial data sets of this sort exist [16.48].

16.7 Comments on Long-Time Evolution Behavior

Once an initial data set satisfying the Einstein constraint equations has been obtained, one can evolve it into a spacetime satisfying the Einstein field equations. As guaranteed by the work discussed in Sect. 16.3, there is a unique globally hyperbolic spacetime development of this initial data which contains (up to diffeomorphism) any other developments of the same set of data.

What do we know about the long-time properties of these maximal developments? The Hawking–Penrose singularity theorems [16.5] tell us that (among spacetimes with a compact Cauchy surface), in one or the other direction in time, such developments *generically* become causally geodesically incomplete, which means that there are causal geodesics in the spacetime which

do not extend to the infinite affine parameter length. This property of causal geodesic incompleteness is, however, consistent with a wide variety of spacetime behavior, including curvature blowup, Cauchy horizon formation, and various topological anomalies [16.53].

One of the intriguing questions concerning spacetime developments is which of these behaviors – curvature blowup, Cauchy horizon formation, or something else – is expected to occur generically among those spacetimes which are causally geodesically incomplete. *Penrose* [16.54] has conjectured that curvature blowup is the generic behavior. This conjecture has been labeled the *strong cosmic censorship* conjecture (SCC) and it is viewed by many as one of the central questions concerning the evolutionary behavior of solutions of Einstein's equations (The strong cosmic censorships conjecture does not imply, and is not implied by, the weak cosmic censorship conjecture, which concerns the generic formation of an event horizon around a singularity which forms in an asymptotically flat solution of Einstein's equations.).

It is well known that there are infinite dimensional families of solutions [16.6] which have bounded curvature and develop Cauchy horizons. The existence of these solutions does *not* disprove SCC. To formulate and study the SCC conjecture carefully, one needs to define the notion of *generic* in terms of the topology of the space of constraint-satisfying initial data sets on a fixed three-dimensional manifold Σ^3, and then determine which sets of initial data evolve into a spacetime with unbounded curvature, and which do not.

Strong cosmic censorship, formulated this way, has been proven for certain families of solutions, most notably (by *Ringstrom* [16.55]) for the Gowdy family which is characterized by the existence of a T^2 isometry group, and vanishing *twists* (The Gowdy family of solutions is introduced and characterized in [16.56] and is studied extensively in [16.57].). Numerical and other formal evidence strongly suggest that it is true for a wider class of spacetimes, with smaller isometry group [16.58]. Proving or disproving this remains a major challenge.

The strong cosmic censorship conjecture concerns generic behavior *among those spacetimes which are causally geodesically incomplete.* Distinct from this issue, and also of very significant interest, is the question of which initial data sets evolve into spacetimes which extend an infinite (proper) time into the future and/or the past, and which do not. A complete answer to this question appears to be beyond our current mathematical capabilities. However, as a small but very significant step towards answering this question, a number of researchers have focussed on the issue of the *stability* – in terms of long time existence and structure – of solutions which exist for infinite proper time.

The landmark stability result in general relativity is the proof by *Christodoulou* and *Klainerman* [16.59] of the stability of Minkowski spacetime. They show, using energies based on the Bel–Robinson tensor to measure initial data perturbations, that the spacetime developments of initial data sets which are sufficiently close to Minkowski initial data do extend an infinite proper time into the future and into the past. Moreover, they show that these developments have the same asymptotic spacetime structure as Minkowski spacetime. These results have been extended to allow electromagnetic as well as gravitational initial data perturbations [16.60] and have also been strengthened in terms of the nature of the asymptotic structure which is shown to be stable [16.60].

If Minkowski spacetime is stable, one might logically proceed to consider if this is also the case for Schwarzschild spacetimes. However, since one knows that a small perturbation of Schwarzschild initial data which adds angular momentum will evolve into a Kerr solution rather than a Schwarzschild solution, one is led to consider the stability of Kerr spacetimes instead. The determination of whether or not Kerr spacetimes are stable is currently one of the most active areas of research in mathematical relativity. A recent report on the research directed towards this goal appears in [16.61].

We note that the stability of other spacetimes has been established: *Friedrichs* [16.62] has shown that De Sitter spacetime is stable, *Andersson* and *Moncrief* [16.63] have shown that Milne spacetime is stable, and *Ringstrom* [16.64] has shown that certain solutions of the Einstein-scalar field equations with accelerating expansion are stable.

References

16.1 R. Arnowitt, S. Deser, C. Misner: The dynamics of gneral relativity. In: *Gravitation: An Introduction to Current Research*, ed. by L. Witten (Wiley, New York 1962) pp. 227–264

16.2 J. Isenberg, J. Nester: Canonical gravity. In: *General Relativity and Gravitation – The Einstein Centenary*, ed. by A. Held (Plenum, New York 1980) pp. 23–93

16.3 Y. Choquet-Bruhat: Théorème d'existence pour certains systèmes d'équations aux dérivées partielles non linéaires, Acta Math. **88**, 141–225 (1952)

16.4 Y. Choquet-Bruhat, R. Geroch: Global aspects of the Cauchy problem in general relativity, Commun. Math. Phys. **14**, 329–335 (1969)

16.5 S. Hawking, G. Ellis: *The Large Scale Structure of Space-Time* (Cambridge Univ. Press, Cambridge 1973)

16.6 V. Moncrief: The space of (generalized) Taub-Nut spacetimes, J. Geom. Phys. **1**, 107–130 (1984)

16.7 H. Ringstrom: *The Cauchy Problem in General Relativity* (European Mathematical Society, Zürich 2009)

16.8 L. Evans: *Partial Differential Equations*, 2nd edn. (AMS, Providence 2010)

16.9 R. Bartnik, J. Isenberg: The constraint equations. In: *The Einstein Equations and the Large Scale Behavior of Gravitational Fields*, ed. by P.T. Chruściel, H. Friedrich (Birkhäuser, Basel 2004) pp. 1–39

16.10 F. Pretorius: Numerical relativity using a generalized harmonic decomposition, Class. Quantum Gravity **22**, 425–452 (2005)

16.11 J. Isenberg, J. Nester: The effect of gravitational interaction on classical fields: A Hamilton Dirac Analysis, Ann. Phys. **107**, 56–81 (1977)

16.12 D. Bao, Y. Choquet-Bruhat, J. Isenberg, P. Yasskin: The well-posedness of ($N = 1$) classical supergravity, J. Math. Phys. **26**, 329–333 (1985)

16.13 A. Fischer, J. Marsden: The Einstein evolution equations as a first-ordeer quasi-linear symmetric hyperbolic system, I, Commun. Math. Phys. **28**, 1–38 (1972)

16.14 T. Hughes, T. Kato, J. Marsden: Well-posed quasi-linear second-order hyperbolic systems with applications to nonlinear elastodynamics and general relativity, Arch. Ration. Mech. Anal. **63**, 273–394 (1977)

16.15 S. Klainerman, I. Rodnianski: Rough solutions to the Einstein vacuum equations, Ann. Math. **61**, 1143–1193 (2005)

16.16 S. Klainerman, I. Rodnianski, J. Szeftel: Overview of the proof of the bounded L^2 curvature conjecture (2012), arXiv:1204.1772v2

16.17 Y. Choquet-Bruhat, J. York: The Cauchy Problem. In: *General Relativity and Gravitation – The Einstein Centenary*, ed. by J. York (Plenum, New York 1980) pp. 99–160

16.18 A. Lichnerowicz: L'integration des equations de la gravitation relativiste et la probleme des *n* corps, Journ. de Math. **XXIII**, 37–63 (1944)

16.19 R. Schoen: Conformal deformation of a Riemannian metric to constant scalar curvature, J. Differ. Geom. **20**, 479–495 (1984)

16.20 J. Isenberg: Constant mean curvature solutions of the Einstein constraint equations on closed manifolds, Class. Quantum Gravity **12**, 2249–2274 (1995)

16.21 J. Barbour, N. O'Murchadha: Conformal superspace: the configuration space of general relativity (2010), arXiv:1009.3559[gr-qc]

16.22 J. Isenberg, D. Maxwell, D. Pollack: A gluing construction for non-vacuum solutions ofthe Einstein constraint equations, Adv. Theor. Math. Phys. **9**, 129–172 (2005)

16.23 Y. Choquet-Bruhat, J. Isenberg, D. Pollack: The constraint equations for the Einstein-scalar field system on compact manifolds, Class. Quantum Gravity **24**, 809–828 (2007)

16.24 E. Hebey, F. Pacard, D. Pollack: A variational analysis of Einstein-scalar field Lichnerowicz equations on compact Riemannian manifolds, Commun. Math. Phys. **278**, 117–132 (2008)

16.25 Y. Choquet-Bruhat, D. Christodoulou: Elliptic systems in $H_{s,\delta}$ spaces on manifolds which are Euclidean at infinity, Acta Math. **146**, 124–150 (1981)

16.26 Y. Choquet-Bruhat, J. Isenberg, J.W. York: Einstein constraints on asymptotically Euclidean manifolds, Phys. Rev. D **61**, 1–20 (2000)

16.27 M. Cantor: The existence of non-trivial asymptotically flat initial data for vacuum spacetimes, Commun. Math. Phys. **57**, 83 (1977)

16.28 D. Brill, M. Cantor: The Laplacian on asymptotically flat manifolds and the specification of scalar curvature, Compos. Math. **43**, 317 (1981)

16.29 D. Maxwell: Solutions of the Einstein constraint equations with apparent horizon boundaries, Commun. Math. Phys. **253**, 561–583 (2004)

16.30 H. Friedrich: Yamabe numbers and the Brill-Cantor criterion, Ann. Henri Poincaré **12**, 1019–1025 (2011)

16.31 L. Andersson, P. Chruściel: Solutions of the constraint equations in general relativity satisfying "hyperboloidal boundary conditions", Diss. Math. (Rozprawy Mat.) **355**, 1–100 (1996)

16.32 P. Chruściel, R. Mazzeo: Initial data sets with ends of cylindrical type I: The Lichnerowicz equation (2012), arXiv:1201.4937

16.33 P. Chruściel, R. Mazzeo, S. Pocchiola: Initial data sets with ends of cylindrical type II: The vector constraint equation arXiv:1201.5138 (2012)

16.34 J. Isenberg, V. Moncrief: A set of nonconstant mean curvature solutions of the Einstein constraint equations on closed manifolds, Class. Quantum Gravity **1**, 1819–1847 (1996)

16.35 P. Allen, A. Clausen, J. Isenberg: Near-constant mean curvature solutions of the Einstein constraint equations with nonnegative Yamabe metrics, Class. Quantum Gravity **25**, 075009 (2008)

16.36 J. Isenberg, N.O. Murchadha: Non-CMC conformal data sets which do not produce solutions of the Einstein constraint equations, Class. Quantum Gravity **21**, S233 (2004)

16.37 M. Dahl, R. Gicquaud, E. Humbert: A limit equation associated to the solvabilityof the vacuum Einstein constraint equations using the conformal method, Duke Math J. **161**, 2669–2697 (2012)

16.38 J. Isenberg, J. Park: Asymptotically hyperbolic non constant mean curvature solutions of the Einstein constraint equations, Class. Quantum Gravity **14**, A189–A202 (1997)

16.39 M. Holst, G. Nagy, G. Tsogtgerel: Rough solutions of the Einstein constraints on closed manifolds without near-CMC conditions, Commun. Math. Phys. **288**, 547–613 (2009)
16.40 D. Maxwell: Rough solutions of the Einstein constraints on compact manifolds, J. Hyperbolic Differ. Equ. **2**, 521–546 (2005)
16.41 R. Gicquaud, A. Sakovich: A large class of non constant mean curvature solutions of the Einstein constraint equations on an asymptotically hyperbolic manifold (2012), arXiv:1012.2246
16.42 D. Maxwell: A model problem or conformal parameterizations of the Einstein constraint equations, Commun. Math. Phys. **302**, 697–736 (2011)
16.43 J.W. York: Conformal "thin-sandwich" data for the initial-value problem of general relativity, Phys. Rev. Lett. **82**, 1350–1353 (1999)
16.44 J. Corvino: Scalar curvature deformation and a gluing construction for the Einstein constraint equations, Commun. Math. Phys. **214**, 137–189 (2000)
16.45 J. Corvino, R. Schoen: On the asymptotics for the vacuum Einstein constraint equations, J. Differ. Geom. **73**, 185–358 (2006)
16.46 P. Chruściel, E. Delay: On mapping properties of the general relativistic constraints operator in weighted function spaces, with applications, Mem. Soc. Math. France **93**, 1–103 (2003)
16.47 J. Isenberg, R. Mazzeo, D. Pollack: Gluing and wormholes for the Einstein constraint equations, Commun. Math. Phys. **231**, 529–568 (2001)
16.48 P. Chruściel, J. Isenberg, D. Pollack: Initial data engineering, Commun. Math. Phys. **257**, 29–42 (2005)
16.49 R. Beig, P. Chruściel, R. Schoen: KIDS are non-generic, Ann. Henri Poincaré **6**, 155–194 (2005)
16.50 P. Chruściel, E. Delay: Existence of non-trivial, vacuum, asymptotically simple spacetimes, Class. Quantum Gravity **19**, L71–L79 (2002)
16.51 P. Chruściel, J. Corvino, J. Isenberg: Construction of n-body initial data sets in general relativity, Commun. Math. Phys. **304**, 637–647 (2011)
16.52 R. Bartnik: Remarks on cosmological spacetimes and constant mean curvature surfaces, Commun. Math. Phys. **117**, 615–624 (1988)
16.53 L. Andersson, T. Barbot, R. Benedetti, F. Bonsante, W. Goldman, F. Labourie, K. Scannell, J. Schlenker: Notes on a paper of Mess (2007), arXiv:0706.0640
16.54 R. Penrose: The question of cosmic censorship, J. Astrophys. Astr. **20**, 233–248 (1999)
16.55 H. Ringstrom: Curvature blow up on a dense subset of the singularity in T^3-Gowdy, J. Hyperbolic Differ. Equ. **2**, 547–564 (2005)
16.56 R. Gowdy: Vacuum spacetimes with two-parameter spacelike isometry groups and compaact invarian hypersurfaces: topologies and boundary conditions, Ann. Phys. **83**, 203–241 (1974)
16.57 P. Chruściel: *On uniqueness in the large of solutions of Einstein's equation (strong cosmic censorship)* (Centre for Mathematics and its Applications, Australian National University 1991)
16.58 J. Isenberg, V. Moncrief: Asymptotic behavior in polarized and half-polarized $U(1)$ symmetric vacuum spacetimes, Class. Quantum Gravity **19**, 5361–5386 (2002)
16.59 D. Christodoulou, S. Klainerman: *The global non linear stability of the Minkowski space* (University Press, Princeton 1993)
16.60 L. Bieri, N. Zipser: Extensions of the Stability Theorem of the Minkowski Space In General Relativity, AMS Studies in Advanced Mathematics (2009)
16.61 M. Dafermos, I. Rodnianski: The black hole stability problem for linear scalar perturbations (2010), arXiv:1201.1797
16.62 H. Friedrich: On the existence of n-geodesically complete or future complete solutions of Einstein's field equations with smooth asymptotic structure, Commun. Math. Phys. **107**, 587–609 (1986)
16.63 L. Andersson, V. Moncrief: *Future complete vacuum spacetimes. In: The Einstein Equations and the Large Scale Behavior of Gravitational Fields* (Birkhäuser, Basel 2004)
16.64 H. Ringstrom: Future stability of the Einstein-non-linear scalar field system, Invent. Math. **173**, 123–208 (2008)

17. Dynamical and Hamiltonian Formulation of General Relativity

Domenico Giulini

Einstein's theory of General Relativity describes spacetime as a solution of a set of non-linear partial differential equations. These equations are initially not in the form of evolution equations and it is hence not clear how to formulate and solve initial-value problems, as would be physically highly desirable. In this contribution it will be shown how to cast Einstein's equations into the form of a constrained Hamiltonian system. This will allow to formulate and solve initial-value problems, integrate Einstein's equations by numerical codes, characterize dynamical degrees of freedom, and characterize isolated systems and their conserved quantities, like energy, momentum, and angular momentum. Moreover, this reformulation of General Relativity is also the starting point for various attempts to subject the gravitational field to the program of canonical quantization. The exposition given here is, to some degree, self contained. It attempts to comprehensively account for all the relevant geometric constructions, including the relevant symplectic geometry of constrained Hamiltonian systems.

17.1 Overview

The purpose of this chapter is to explain how the field equations of general relativity – often simply referred to as *Einstein's equations* – can be understood as a dynamical system; more precisely, as a *constrained Hamiltonian system.*

In general relativity, it is often said, spacetime becomes dynamical. This is meant to say that the geometric structure of spacetime is encoded in a field that, in turn, is subject to local laws of propagation and coupling, just as, e.g., the electromagnetic field. It is *not* meant to say that spacetime as a whole evolves. Spacetime does not evolve, spacetime just is. But a given spacetime (four dimensional) can be viewed as the evolution, or history, of space (three dimensional). There is

a huge redundancy in this representation, in the sense that apparently very different evolutions of space represent the same spacetime. However, if the resulting spacetime is to satisfy Einstein's equations, the evolution of space must also obey certain well-defined restrictions. Hence, the task is to give precise mathematical expression to the redundancies in representation as well as the restrictions of evolution for this picture of spacetime as space's history. This will be the main task of this chapter.

This dynamical picture will be important for posing and solving time-dependent problems in general relativity (GR), like the scattering of black holes with its subsequent generation and radiation of gravitational waves. It is also a key technology to:

- Formulate and solve initial-value problems.
- Integrate Einstein's equations by numerical codes.
- Characterize dynamical degrees of freedom.
- Characterize isolated systems and the association of asymptotic symmetry groups, which will give rise to globally conserved *charges*, like energy and linear as well as angular momenta (Poincaré charges).

Moreover, it is also the starting point for the *canonical quantization program*, which constitutes one main approach to the yet unsolved problem of *quantum gravity*. In this approach one tries to make essential use of the Hamiltonian structure of the classical theory in formulating the corresponding quantum theory. This strategy has been applied successfully in the transition from classical to quantum mechanics and also in the transition from classical to quantum electrodynamics. Hence, the canonical approach to quantum gravity may be regarded as conservative, insofar as it tries to apply otherwise established rules to a classical theory that is experimentally and observationally extremely well tested. The underlying hypothesis here is that we may quantize interaction-wise. This distinguishes this approach from string theory, the underlying credo of which is that quantum gravity only makes sense on the basis of a unified description of all interactions.

Historically the first paper to address the problem of how to put Einstein's equations into the form of a Hamiltonian dynamical system was that of *Dirac* [17.1] in 1958. He also noticed its constrained nature and started to develop the corresponding generalization of *constrained Hamiltonian systems* in [17.2] and their quantization [17.3]. On the classical side, this developed into the more geometric *Dirac–Bergmann theory* of constraints [17.4] and on the quantum side into an elaborate theory of quantization of systems with gauge redundancies; see [17.5] for a comprehensive account. Dirac's attempts were soon complemented by an extensive joint work of Richard Arnowitt, Stanley Deser, and Charles Misner – usually and henceforth abbreviated by ADM. Their work started in 1959 by a paper [17.6] of the first two of these authors and continued in the series [17.7–18] of 12 more papers by all three. A comprehensive summary of their work was given in 1962 in [17.19], which was republished in 2008 in [17.20]; see also the editorial note [17.21] with short biographies of ADM. Modern textbooks on the $3+1$ formalism and the canonical formulation of GR, with applications to cosmology, black holes, and quantum gravity, are [17.22, 23], respectively.

17.2 Notation and Conventions

Throughout, GR stands for general relativity. Spacetime is a differentiable manifold M of dimension n, endowed with a metric g of signature $(\varepsilon, +, \dots, +)$. In GR $n=4$ and $\varepsilon=-1$ and it is implicitly understood that these are the *right* values. However, either for the sake of generality and/or particular interest, we will sometimes state formulae for general n and ε, where usually $n \geq 2$ (sometimes $n \geq 3$) and either $\varepsilon=-1$ (Lorentzian metric) or $\varepsilon=+1$ (Riemannian metric; also called Euclidean metric). The case $\varepsilon=1$ has been extensively considered in path-integral approaches to quantum gravity, there referred to as *Euclidean quantum gravity*.

The tangent space of M at point $p \in M$ will be denoted by T_pM, the cotangent space by T_p^*M, and the tensor product of u factors T_pM with d factors of T_p^*M by $T^u_{pd}M$. (Mnemonic in components: $u=$ number of indices *upstairs*, $d=$ number of indices *downstairs*.) An element T in $T^u_{pd}M$ is called a tensor of contravariant rank u and covariant rank d at point p, or simply a tensor of rank (u,d) at p. T is called *contravariant* if $d=0$ and $u>0$, and *covariant* if $u=0$ and $d>0$. A tensor with $u>0$ and $d>0$ is then referred to as of *mixed type*. Note that $T_pM = T^1_{p0}M$ and $T_p^*M = T^0_{p1}M$. The set of tensor *fields*, i. e. smooth assignments of an element

in $T^u_{pd}M$ for each $p \in M$, is denoted by $\Gamma T^u_d M$. Unless stated otherwise, *smooth* means C^∞, i. e. continuously differentiable to any order. For $t \in \Gamma T^u_d M$, we denote by $t_p \in T^u_{pd}M$ the evaluation of t at $p \in M$. $C^\infty(M)$ denotes the set of all C^∞ real-valued functions on M, which we often simply call smooth functions.

Note that the general definition of *metric* is as follows: $g \in \Gamma T^0_2 M$, such that g_p is a symmetric nondegenerate bilinear form on T_pM. Such a metric provides isomorphisms (sometimes called the *musical isomorphisms*)

$$\begin{aligned} \flat : T_pM &\to T^*_pM \\ X &\mapsto X^\flat := g(X, \cdot)\,, \end{aligned} \tag{17.1a}$$

$$\begin{aligned} \sharp : T^*_pM &\to T_pM \\ \omega &\mapsto \omega^\sharp := \flat^{-1}(\omega)\,. \end{aligned} \tag{17.1b}$$

Using $\sharp$, we obtain a metric g_p^{-1} on T^*_pM from the metric g_p on T_pM as follows

$$g_p^{-1}(\omega_1, \omega_2) := g_p(\omega_1^\sharp, \omega_2^\sharp) = \omega_1(\omega_2^\sharp)\,. \tag{17.2}$$

We also recall that the tensor space $T^1_{p1}M$ is naturally isomorphic to the linear space $\mathrm{End}(T_pM)$ of all endomorphisms (linear self maps) of T_pM. Hence, it carries a natural structure as associative algebra, the product being composition of maps denoted by $\circ$. As usual, the *trace*, denoted Tr, and the *determinant*, denoted det, are the naturally defined real-valued functions on the space of endomorphisms. For purely co- or contravariant tensors the trace can be defined by first applying one of the isomorphisms (17.1). In this case we write Tr_g to indicate the dependence on the metric g.

Geometric representatives of curvature are often denoted by bold-faced abbreviations of their names, like **Riem** and **Weyl** for the (covariant, i. e. all indices down) Riemann and Weyl tensors, **Sec** for the sectional curvature, **Ric** and **Ein** for the Ricci and Einstein tensors, **Scal** for the scalar curvature, and **Wein** for the Weingarten map (which is essentially equivalent to the extrinsic curvature). This is done in order to highlight the geometric meaning behind some basic formulae, at least the simpler ones. Later, as algebraic expressions become more involved, we will also employ the standard component notation for computational ease.

17.3 Einstein's Equations

Einstein's equations form a set of 10 quasi-linear partial differential equations of second order in spacetime. At each point of spacetime (event) they equate 10 purely geometric quantities to 10 quantities encoding the densities (quantity per unit volume) and current densities (quantity per unit area and unit time) of energy and momentum of matter. The geometric quantity in Einstein's equations is the Einstein tensor **Ein**; the matter quantity is the energy–momentum tensor **T**. Both tensors are of second rank, symmetric, and here taken to be covariant (in components: *all indices down*).

Einstein's equations (actually a single tensor equation, but throughout we use the plural to emphasize that it comprises several component equations) state the simple proportionality of **Ein** with **T**

$$\mathbf{Ein} = \kappa \mathbf{T}\,, \tag{17.3}$$

where κ denotes the dimensionful constant of proportionality. Note that no explicit reference to the dimension n enters (17.3), so that even if $n \neq 4$ it is usually referred to as Einstein's equations. We could have explicitly added a *cosmological constant* term $g\Lambda$ on the left-hand side, where Λ is a constant the physical dimension of which is the square of an inverse length. However, as long as we write down our formulae for general **T** we may absorb this term into **T**, where it accounts for a contribution $\mathbf{T}_\Lambda = -g\Lambda/\kappa$. This has to be kept in mind when explicit models for **T** are used and when we speak of *vacuum*, which now means

$$\mathbf{T}_{\text{vacuum}} = \mathbf{T}_\Lambda := -\kappa^{-1} g\Lambda\,. \tag{17.4}$$

The signs are chosen such that a positive Λ accounts for a positive energy density and a negative pressure if the spacetime is Lorentzian ($\varepsilon = -1$).

There is another form of Einstein's equations which is sometimes advantageous to use and in which n explicitly enters

$$\mathbf{Ric} = \kappa \left(\mathbf{T} - \frac{1}{n-2} g \mathrm{Tr}_g(\mathbf{T}) \right)\,. \tag{17.5}$$

These two forms are easily seen to be mathematically equivalent by the identities

$$\mathbf{Ein} = \mathbf{Ric} - \frac{1}{2} g \mathrm{Tr}_g(\mathbf{Ric}) \,, \tag{17.6a}$$

$$\mathbf{Ric} = \mathbf{Ein} - \frac{1}{n-2} g \mathrm{Tr}_g(\mathbf{Ein}) \,. \tag{17.6b}$$

With respect to a local field of basis vectors $\{e_0, e_1, \ldots, e_{n-1}\}$, we write $\mathbf{Ein}(e_\mu, e_\nu) =: G_{\mu\nu}$, $\mathbf{T}(e_\mu, e_\nu) =: T_{\mu\nu}$, and $\mathbf{Ric}(e_\mu, e_\nu) =: R_{\mu\nu}$. Then (17.3) and (17.5) take on the component forms

$$G_{\mu\nu} = \kappa T_{\mu\nu} \tag{17.7}$$

and

$$R_{\mu\nu} = \kappa \left(T_{\mu\nu} - \frac{1}{n-2} g_{\mu\nu} T^\lambda_\lambda \right) , \tag{17.8}$$

respectively. Next we explain the meanings of the symbols in Einstein's equations from left to right.

17.3.1 What Aspects of Geometry?

The left-hand side of Einstein's equations comprises certain measures of curvature. As will be explained in detail in Sect. 17.5, all curvature information in dimensions higher than two can be reduced to that of *sectional curvature*. The sectional curvature at a point $p \in M$ tangent to $\mathrm{span}\{X, Y\} \subset T_pM$ is the Gauss curvature at p of the submanifold spanned by the geodesics in M emanating from p tangent to $\mathrm{span}\{X, Y\}$. The Gauss curvature is defined as the product of two principal curvatures, each being measured in units of an inverse length (the inverse of a principal radius). Hence, the Gauss curvature is measured in units of an inverse length squared.

At each point p in spacetime the Einstein and Ricci tensors are symmetric bilinear forms on T_pM. Hence, $\mathbf{Ein}_p$ and $\mathbf{Ric}_p$ are determined by the values $\mathbf{Ein}_p(W, W)$ and $\mathbf{Ric}_p(W, W)$ for all $W \in T_pM$. By continuity in W, this remains true if we restrict W to the open and dense set of vectors which are not null, i. e. for which $g(W, W) \neq 0$. As we will see later on, we then have

$$\mathbf{Ein}(W, W) = -g(W, W) \sum_{\perp W}^{N_1} \mathbf{Sec} \,, \tag{17.9}$$

$$\mathbf{Ric}(W, W) = +g(W, W) \sum_{\parallel W}^{N_2} \mathbf{Sec} \,. \tag{17.10}$$

For the Einstein tensor the sum on the right-hand side is over any complete set of $N_1 = \frac{1}{2}(n-1)(n-2)$ sectional curvatures of pairwise-orthogonal planes in the orthogonal complement of W in T_pM. For the Ricci tensor it is over any complete set of $N_2 = n - 1$ sectional curvatures of pairwise-orthogonal planes containing W. If W is a timelike unit vector representing an observer, $\mathbf{Ein}(W, W)$ is simply $(-\varepsilon)$ times an equally weighted sum of spacelike sectional curvatures, whereas $\mathbf{Ric}(W, W)$ is ε times an equally weighted sum of timelike sectional curvatures. In that sense we may say that, e.g., $\mathbf{Ein}(W, W)$ at $p \in M$ measures the mean Gauss curvature of the (local) hypersurface in M that is spanned by geodesics emanating from W orthogonal to W. It, too, is measured in units of the square of an inverse length.

17.3.2 What Aspects of Matter?

Now we turn to the right-hand side of Einstein's equations. We restrict to four spacetime dimensions, though much of what we say will apply verbatim to other dimensions. The tensor $\mathbf{T}$ on the right-hand side of (17.3) is the *energy–momentum tensor* of matter. With respect to an orthonormal basis $\{e_0, e_1, \ldots, e_{n-1}\}$ with timelike e_0, the components $T_{\mu\nu} := \mathbf{T}(e_\mu, e_\nu)$ form a symmetric 4×4 matrix, which we represent as follows by splitting off terms involving a time component

$$T_{\mu\nu} = \begin{pmatrix} \varepsilon & -c\vec{\mathcal{M}} \\ -\frac{1}{c}\vec{S} & T_{mn} \end{pmatrix} . \tag{17.11}$$

Here all matrix elements refer to the matter's energy–momentum distribution relative to the rest frame of the observer who momentarily moves along e_0 (i. e. with four-velocity $u = ce_0$) and uses the basis $\{e_1, e_2, e_3\}$ in his/her rest frame. Then $\varepsilon = T_{00}$ is the energy density, $\vec{S} = (s_1, s_2, s_3)$ the (components of the) energy current density, i. e. energy per unit surface area and unit time interval, $\vec{\mathcal{M}}$ the momentum density, and finally T_{mn} the (components of the) momentum current density, i. e. momentum per unit of area and unit time interval. Note that the symmetry $T_{\mu\nu} = T_{\nu\mu}$ implies a simple relation between the energy current density and the momentum density

$$\vec{S} = c^2 \vec{\mathcal{M}} \,. \tag{17.12}$$

The remaining relations $T_{mn} = T_{nm}$ express equality of the m-th component of the current density for n-momentum with the n-th component of the current

density for m-momentum. Note that the two minus signs in front of the mixed components of (17.11) would have disappeared had we written down the contravariant components $T^{\mu\nu}$. In flat spacetime, the four equations $\partial^\mu T_{\mu\nu}$ express the local conservation of energy and momentum. In curved spacetime, the identity (compare Sect. 17.5)

$$\nabla^\mu G_{\mu\nu} \equiv 0 \tag{17.13}$$

implies via (17.7)

$$\nabla^\mu T_{\mu\nu} = 0\,, \tag{17.14}$$

which may be interpreted as expressing a local conservation of energy and momentum for the matter *plus* the gravitational field, though there is no such thing as a separate energy–momentum tensor on spacetime for the gravitational field.

Several positivity conditions can be imposed on the energy–momentum tensor $\mathbf{T}$. The simplest is known as the *weak energy condition* and reads $\mathbf{T}(W, W) \geq 0$ for all timelike W. It is equivalent to the requirement that the energy density measured by any local observer is nonnegative. For a perfect fluid of rest-mass density ρ and pressure p, the weak energy condition is equivalent to both conditions $\rho \geq 0$ and $p \geq -c^2\rho$. The strong energy condition says that $\left(\mathbf{T} - \frac{1}{2}g\mathrm{Tr}_g(\mathbf{T})\right)(W, W) \geq 0$ again for all timelike W. This neither follows nor implies the weak energy condition. For a perfect fluid, it is equivalent to both conditions $p \geq -c^2\rho$ and $p \geq -c^2\rho/3$, i. e. to the latter alone if ρ is positive and to the former alone if ρ is negative (which is not excluded here). Its significance lies in the fact that it ensures attractivity of gravity as described by Einstein's equations. It must, for example, be violated if matter is to drive inflation. Note that upon imposing Einstein's equations the weak and the strong energy conditions read $\mathbf{Ein}(W, W) \geq 0$ and $\mathbf{Ric}(W, W) \geq 0$, respectively. From (17.9) and (17.10), we can see that for fixed W these imply conditions on complementary sets of sectional curvatures.

17.3.3 A Small Digression on Symmetries

Conservation laws for the matter alone result in the presence of symmetries. If V is a Killing vector field, i. e. one has $L_V g = 0$, where L_V denotes the Lie derivative with respect to V, the 1-form J_V that results from contracting $\mathbf{T}$ with V is divergence free

$$J_V := i_V\mathbf{T} = V^\mu T_{\mu\nu}\,\mathrm{d}x^\nu\,, \tag{17.15}$$

where, because of Killing's equation,

$$L_V g = 0 \Leftrightarrow \nabla_\mu V_\nu + \nabla_\nu V_\mu = 0\,, \tag{17.16}$$

and (17.14) one has

$$\nabla_\mu J^\mu_V = 0\,. \tag{17.17}$$

This may be equivalently expressed by saying that the 3-form $\star J_V$, which is the Hodge dual of the 1-form J_V, is closed

$$\mathrm{d}\star J_V = 0\,. \tag{17.18}$$

Integrating $\star J_V$ over some three-dimensional submanifold Σ results in a quantity

$$Q[V, \Sigma] := \int_\Sigma \star J_V\,, \tag{17.19}$$

which, because of (17.18), is largely independent of Σ. More precisely, if $\Omega \subset M$ is an oriented domain with boundary $\partial\Omega = \Sigma_1 - \Sigma_2$, then Stokes' theorem gives $Q[V, \Sigma_1] = Q[V, \Sigma_2]$. Suppose now that V arises from a finite-dimensional Lie group G that acts from the left on (M, g) by isometries. Then this defines an anti-homomorphism from the Lie algebra $\mathrm{Lie}(G)$ of G into the Lie algebra of vector fields on M. (For a right action we would have obtained an ordinary homomorphism, but this is not important here.) This we denote by $V : \xi \mapsto V_\xi$ for $\xi \in \mathrm{Lie}(G)$, so that $[V_\xi, V_\eta] = -V_{[\xi,\eta]}$. For fixed Σ the integral (17.18) then becomes a linear map from $\mathrm{Lie}(G)$ to $\mathbb{R}$

$$P : \mathrm{Lie}(G) \to \mathbb{R}\,, \quad P(\xi) := Q[V_\xi, \Sigma]\,. \tag{17.20}$$

Hence, each hypersurface Σ defines an element $P \in \mathrm{Lie}^*(G)$ in the vector space that is dual to the Lie algebra, given that the integral over Σ converges. This is the case for spacelike Σ and energy–momentum tensors with spatially compact support (or at least sufficiently rapid fall off). The same argument as above using Stokes' theorem and (17.18) then shows that P is independent of the choice of spacelike Σ. In other words, we obtain a conserved quantity $P \in \mathrm{Lie}^*(G)$ for G-symmetric spacetimes satisfying Einstein's equations. This map may be called the *momentum map*. (Compare the notion of a momentum map in Hamiltonian mechanics; cf. Sect. 17.7.) Note that the value of the momentum map is a quantity that is globally associated with all of spacetime, not a particular region or

point of it. The value lies in the vector space Lie*(G) which carries the co-adjoint representation, Ad*, of G. One may show that under the assumption of suitable covariance properties for **T**, the momentum map, considered as a map from the matter fields to Lie*(G), is Ad* equivariant.

17.3.4 How Do Geometry and Matter Relate Quantitatively?

We return to Einstein's equations and finally discuss the constant of proportionality κ on the right-hand side of (17.3). It is given by (J = joule)

$$\kappa := \frac{8\pi G}{c^4} \approx 2.1 \times 10^{-43}\ \frac{\mathrm{m}^{-2}}{\mathrm{J\,m}^{-3}}\,, \tag{17.21}$$

where $G \approx 6.67384(80) \times 10^{-11}\,\mathrm{m}^3\,\mathrm{kg}^{-1}\,\mathrm{s}^{-2}$ is Newton's constant. It is currently (March 2013) known with a relative standard uncertainty of 1.2×10^{-4} and is thus by far the least well known of the fundamental physical constants. $c = 299\,792\,458\,\mathrm{m\,s}^{-1}$ is the vacuum speed of light whose value is exact, due to the SI definition of the meter (*the meter is the length of the path traveled by light in vacuum during a time interval of* 1/299 792 458 *of a second*).

The physical dimension of κ is $\mathrm{T}^2/(\mathrm{M}\cdot\mathrm{L})$, that is, in SI units, $s^2 \cdot kg^{-1} \cdot m^{-1}$ or $\mathrm{m}^{-2}/(\mathrm{J\ m}^{-3})$, where $\mathrm{J} = \text{joule} = \mathrm{kg\,m}^2\,\mathrm{s}^{-2}$. It converts the common physical dimension of all components $T_{\mu\nu}$, which is that of an energy density (joule per cubic meter in SI units) into that of the components of **Ein**, which is that of curvature (in dimension ≥ 2), i. e., the square of an inverse length (inverse square meter in SI units).

If we express energy density as mass density times c^2, the conversion factor is $\kappa c^2 = 8\pi G/c^2$. It can be expressed in various units that give a feel for the local *curving power* of mass densities. For that of water, $\rho_\mathrm{W} \approx 10^3\,\mathrm{kg\,m}^{-3}$, and nuclear matter in the core of a neutron star (which is more than twice that of atomic nuclei), $\rho_\mathrm{N} \approx 5 \times 10^{17}\,\mathrm{kg\,m}^{-3}$, we get, respectively

$$\kappa c^2 \approx \left(\frac{1}{1.5\,\mathrm{AU}}\right)^2 \cdot \rho_\mathrm{W}^{-1} \approx \left(\frac{1}{10\,\mathrm{km}}\right)^2 \cdot \rho_\mathrm{N}^{-1}\,, \tag{17.22}$$

where $\mathrm{AU} = 1.5 \times 10^{11}\,\mathrm{m}$ is the astronomical unit (mean Earth–Sun distance). Hence, roughly speaking, matter densities of water cause curvature radii of the order of the astronomical unit, whereas the highest known densities of nuclear matter cause curvature radii of tens of kilometers. The curvature caused by mere mass density is that expressed in $\mathbf{Ein}(W, W)$ when W is taken to be the unit timelike vector characterizing the local rest frame of the matter: it is a mean of spatial sectional curvatures in the matter's local rest frame. Analogous interpretations can be given for the curvatures caused by momentum densities (energy current densities) and momentum current densities (stresses).

17.4 Spacetime Decomposition

In this section we explain how to decompose a given spacetime (M, g) into *space* and *time*. For this to be possible we need to make the assumption that M is diffeomorphic to the product of the real line $\mathbb{R}$ and some 3-manifold Σ

$$M \cong \mathbb{R} \times \Sigma\,. \tag{17.23}$$

This will necessarily be the case for *globally hyperbolic* spacetimes, i. e. spacetimes admitting a *Cauchy surface* [17.24]. We assume Σ to be orientable, for, if it were not, we could take the orientable double cover of it instead. Orientable 3-manifolds are always parallelizable [17.25], i. e. admit three globally defined and pointwise linearly independent vector fields. This is equivalent to the triviality of the tangent bundle. In the closed case this is known as *Stiefel's theorem* (compare [17.26], problem 12-B) and in the open case it follows, e.g., from the well-known fact that every open 3-manifold can be immersed in $\mathbb{R}^3$ [17.27]. Note that orientability is truly necessary; e.g., $\mathbb{RP}^2 \times S^1$ is not parallelizable. Since Cartesian products of parallelizable manifolds are again parallelizable, it follows that a four-dimensional product spacetime (17.23) is also parallelizable. This does, of course, not generalize to higher dimensions. Now, for noncompact four-dimensional spacetimes it is known from [17.28] that parallelizability is equivalent to the existence of a *spin structure*, without which spinor fields could not be defined on spacetime (compare Chap. 15). So, we see that the existence of spin structure is already implied by (17.23) and hence does not pose any further

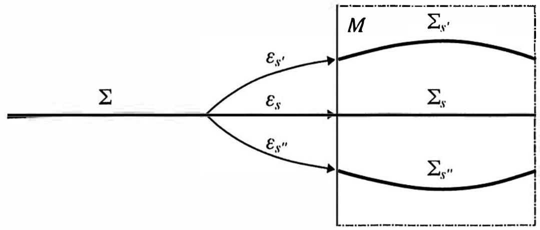


Fig. 17.1 Spacetime M is foliated by a one-parameter family of spacelike embeddings of the 3-manifold Σ. Here the image $\Sigma_{s'}$ of Σ under $\varepsilon_{s'}$ lies to the future (*above*) and $\Sigma_{s''}$ to the past (*below*) of Σ_s if $s'' < s < s'$. *Future* and *past* refer to the time function t, which has so far not been given any metric significance

topological restriction. Note that the only other potential topological restriction at this stage is that imposed from the requirement that a smooth Lorentz metric is to exist everywhere on spacetime. This is equivalent to a vanishing Euler characteristic (see, e.g., [17.25, §40]), which in turn is equivalent to the global existence of a continuous, nowhere-vanishing vector field (possibly up to sign) on spacetime. But such a vector field clearly exists on any Cartesian product with one factor being $\mathbb{R}$. We conclude that existence of a Lorentz metric and a spin structure on an orientable spacetime $M = \mathbb{R} \times \Sigma$ poses no restrictions on the topology of an orientable Σ. As we will see later on, even Einstein's equations pose *no* topological restriction on Σ, in the sense that *some* (physically reasonable) solutions to Einstein's equations exist for any given Σ. Topological restrictions may occur, however, if we ask for solutions with special properties (see below).

Now, given Σ, we consider a one-parameter family of embeddings (see Fig. 17.1)

$$\varepsilon_s : \Sigma \to M\,, \quad \Sigma_s := \varepsilon_s(\Sigma) \subset M\,. \tag{17.24}$$

We distinguish between the abstract 3-manifold Σ and its image Σ_s in M. The latter is called the leaf corresponding to the value $s \in \mathbb{R}$. Each point in M is contained in precisely one leaf. Hence, there is a real-valued function $t : M \to \mathbb{R}$ that assigns to each point in M the parameter value of the leaf it lies on

$$t(p) = s \Leftrightarrow p \in \Sigma_s\,. \tag{17.25}$$

So far this is only a foliation of spacetime by three-dimensional leaves. For them to be addressed as *space*, the metric induced on them must be positive definite, that is, the leaves should be spacelike submanifolds. This means that the 1-form $\mathrm{d}t$ is timelike

$$g^{-1}(\mathrm{d}t, \mathrm{d}t) < 0\,. \tag{17.26}$$

The normalized field of 1-forms is then

$$n^\flat := \frac{\mathrm{d}t}{\sqrt{-g^{-1}(\mathrm{d}t, \mathrm{d}t)}}\,. \tag{17.27}$$

As explained in Sect. 17.2, we write $n^\flat$ since we think of this 1-form as the image under g of the normalized vector field perpendicular to the leaves

$$n^\flat = g(n, \cdot)\,. \tag{17.28}$$

The linear subspace of vectors in T_pM which are tangent to the leaf through p is denoted by $T_p^\parallel M$; hence

$$T_p^\parallel M := \{X \in T_pM \mid \mathrm{d}t(X) = 0\}\,. \tag{17.29}$$

The orthogonal complement is just the span of n at p, which we denote by $T_p^\perp M$. This gives, at each point p of M, the g-orthogonal direct sum

$$T_pM = T_p^\perp M \oplus T_p^\parallel M \tag{17.30}$$

and associated projections (we drop reference to the point p)

$$\begin{aligned} P^\perp :\ & TM \to T^\perp M\,, \\ & X \mapsto \varepsilon g(X, n)\, n\,, \end{aligned} \tag{17.31a}$$

$$\begin{aligned} P^\parallel :\ & TM \to T^\parallel M\,, \\ & X \mapsto X - \varepsilon g(X, n)\, n\,. \end{aligned} \tag{17.31b}$$

As already announced in Sect. 17.2, we introduced the symbol

$$\varepsilon = g(n, n)\,, \tag{17.32}$$

in order to keep track of where the signature matters. Note that the projection operators (17.31) are self adjoint with respect to g, so that for all X, $Y \in TM$ we have

$$g(P^\perp X, Y) = g(X, P^\perp Y), \tag{17.33a}$$

$$g(P^\parallel X, Y) = g(X, P^\parallel Y)\,. \tag{17.33b}$$

A vector is called *horizontal* iff it is in the kernel of $P^\perp$, which is equivalent to being invariant under $P^\parallel$.

It is called *vertical* iff it is in the kernel of $P^{\|}$, which is equivalent to being invariant under $P^{\perp}$.

All this can be extended to forms. We define *vertical* and *horizontal forms* as those annihilating horizontal and vertical vectors, respectively

$$T_p^{*\perp}M := \{\omega \in T_p^*M \mid \omega(X) = 0\,,\ \forall X \in T_p^{\|}M\}\,, \tag{17.34a}$$

$$T_p^{*\|}M := \{\omega \in T_p^*M \mid \omega(X) = 0\,,\ \forall X \in T_p^{\perp}M\}\,. \tag{17.34b}$$

Using the *musical* isomorphisms (17.1), the self-adjoint projection maps (17.31) on vectors define self-adjoint projection maps on covectors (again dropping the reference to the base point p)

$$P_*^{\perp} := \flat \circ P^{\perp} \circ \sharp \,:\, T^*M \to T^{*\perp}M\,, \tag{17.35a}$$

$$P_*^{\|} := \flat \circ P^{\|} \circ \sharp \,:\, T^*M \to T^{*\|}M\,. \tag{17.35b}$$

For example, letting the horizontal projection of the form ω act on the vector X, we get

$$\begin{aligned} P_*^{\|}\omega(X) &= (P^{\|}\omega^{\sharp})^{\flat}(X) \\ &= g(P^{\|}\omega^{\sharp}, X) \\ &= g(\omega^{\sharp}, P^{\|}X) \\ &= \omega(P^{\|}X)\,, \end{aligned} \tag{17.36}$$

where we merely used the definitions (17.1) of $\flat$ and $\sharp$ in the second and fourth equalities, respectively, and the self adjointness (17.33b) of $P^{\|}$ in the third equality. The analogous relation holds for $P_*^{\perp}\omega(X)$. It is also straightforward to check that $P_*^{\|}$ and $P_*^{\perp}$ are self adjoint with respect to g^{-1} (cf. (17.2)).

Having the projections defined for vectors and covectors, we can also define them for the whole tensor algebra of the underlying vector space, just by taking the appropriate tensor products of these maps. All tensor products between $P^{\|}$ and $P_*^{\|}$ will then, for simplicity, just be denoted by $P^{\|}$, the action on the tensor being obvious. Similarly for $P^{\perp}$. (In what follows we need not consider mixed projections.) The projections being pointwise operations, we can now define vertical and horizontal projections of arbitrary tensor fields. Hence, a tensor field $T \in \Gamma T_d^uM$ is called *horizontal* iff $P^{\|}T = T$. The space of horizontal tensor fields of rank (u,d) is denoted by $\Gamma T^{\|}{}_d^uM$.

As an example, the horizontal projection of the metric g is

$$h := P^{\|}g := g(P^{\|}\cdot\,, P^{\|}\cdot) = g - \varepsilon n^{\flat} \otimes n^{\flat}\,. \tag{17.37}$$

Hence, $h \in \Gamma T^{\|}{}_2^0M$. Another example of a horizontal vector field is the *acceleration* of the normal field n

$$a := \nabla_n n\,. \tag{17.38}$$

Here ∇ denotes the Levi-Civita covariant derivative with respect to g. An observer who moves perpendicular to the horizontal leaves has four-velocity $u = cn$ and four-acceleration c^2a. If L denotes the Lie derivative, it is easy to show that the acceleration 1-form satisfies

$$a^{\flat} = L_n n^{\flat}\,. \tag{17.39}$$

Moreover, as n is hypersurface orthogonal, it is irrotational; hence, its 1-form equivalent satisfies

$$\mathrm{d}n^{\flat} \wedge n^{\flat} = 0\,, \tag{17.40a}$$

which is equivalent to the condition of vanishing horizontal curl

$$P^{\|}\,\mathrm{d}n^{\flat} = 0\,. \tag{17.40b}$$

Equation (17.40a) can also be immediately inferred directly from (17.27). Taking the operation $i_n \circ d$ (exterior derivative followed by contraction with n) as well as the Lie derivative with respect to n of (17.39) shows that

$$\mathrm{d}a^{\flat} \wedge n^{\flat} = 0\,, \tag{17.41a}$$

an equivalent expression being again the vanishing of the horizontal curl of a

$$P^{\|}\,\mathrm{d}a^{\flat} = 0\,. \tag{17.41b}$$

This will be useful later on.

Note that a is a horizontal covector field, i. e. an element of $\Gamma T^{\|}{}_{d=1}^{u=0}M$. More generally, for a purely covariant horizontal tensor field we have the following results, which will also be useful later on: let $T \in \Gamma T^{\|}{}_d^0M$; then

$$P^{\|}L_nT = L_nT\,, \tag{17.42a}$$

$$L_{fn}T = fL_nT\,, \tag{17.42b}$$

for all $f \in C^{\infty}(M)$. Note that (17.42a) states that the Lie derivative in the normal direction of a horizontal covariant tensor field is again horizontal. That this is not entirely evident follows, e.g., from the fact that a corresponding result does not hold for $T \in \Gamma T^{\|}{}_d^uM$ where

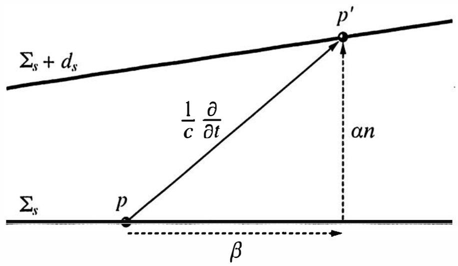


Fig. 17.2 For fixed $q \in \Sigma$ its image points $p = \varepsilon_s(q)$ and $p' = \varepsilon_{s+\mathrm{d}s}(q)$ for infinitesimal $\mathrm{d}s$ are connected by the vector $\partial/\partial t|_p$, whose components normal to Σ_s are α (one function, called lapse) and β (three functions, called shift), respectively

$u > 0$. The proofs of (17.42) just use standard manipulations.

A fixed space point $q \in \Sigma$ defines the worldline (history of that point) $\mathbb{R} \ni s \mapsto \varepsilon_s(q)$. The collection of all worldlines of all space points defines a foliation of M into one-dimensional timelike leaves. Each leaf is now labeled uniquely by a space point. We can think of *space*, i. e., the abstract manifold Σ, as the quotient $M/\sim$, where $p \sim p'$ iff both points lie on the same worldline. As any Σ_s intersects each worldline exactly once, each Σ_s is a representative of space. Instead of using the foliation by three-dimensional spatial leaves (17.24), we could have started with a foliation by timelike lines, plus the condition that these lines are vorticity free. These two concepts are equivalent. Depending on the context, one might prefer to emphasize one or the other.

The vector parallel to the worldline at $p = \varepsilon_s(q)$ is, as usual in differential geometry, defined by its action on $f \in C^\infty(M)$ (smooth, real-valued functions)

$$\left.\frac{\partial}{\partial t}\right|_{\varepsilon_s(q)} f = \left.\frac{\mathrm{d}f\left(\varepsilon_{s'}(q)\right)}{\mathrm{d}s'}\right|_{s'=s} . \tag{17.43}$$

At each point this vector field can be decomposed into its horizontal component that is tangential to the leaves of the given foliation and its normal component. We write

$$\frac{1}{c}\frac{\partial}{\partial t} = \alpha n + \beta , \tag{17.44}$$

where β is the tangential part; see Fig. 17.2.

The real-valued function α is called the *lapse* (function) and the horizontal vector field β is called the *shift* (vector field).

17.4.1 Decomposition of the Metric

Let $\{e_0, e_1, e_2, e_3\}$ be a locally defined orthonormal frame with dual frame $\{\theta^0, \theta^1, \theta^2, \theta^3\}$. We call them *adapted* to the foliation if $e_0 = n$ and $\theta^0 = n^\flat$. A local coordinate system $\{x^0, x^1, x^2, x^3\}$ is called *adapted* if $\partial/\partial x^a$ are horizontal for $a = 1, 2, 3$. Note that in the latter case $\partial/\partial x^0$ is not required to be orthogonal to the leaves (i. e. it need not be parallel to n). For example, we may take x^0 to be proportional to t; say $x^0 = ct$.

In the orthonormal coframe the spacetime metric, i. e. the field of signature $(\varepsilon, +, +, +)$ metrics in the tangent spaces, has the simple form

$$g = \varepsilon\theta^0 \otimes \theta^0 + \sum_{a=1}^{3} \theta^a \otimes \theta^a . \tag{17.45}$$

The inverse spacetime metric, i. e. the field of signature $(\varepsilon, +, +, +)$ metrics in the cotangent spaces, has the form

$$g^{-1} = \varepsilon e_0 \otimes e_0 + \sum_{a=1}^{3} e_a \otimes e_a . \tag{17.46}$$

The relation that expresses the coordinate basis in terms of the orthonormal basis is of the form (in a self-explanatory matrix notation)

$$\begin{pmatrix} \frac{\partial}{\partial x^0} \\ \frac{\partial}{\partial x^m} \end{pmatrix} = \begin{pmatrix} \alpha & \beta^a \\ 0 & A^a_m \end{pmatrix} \begin{pmatrix} e_0 \\ e_a \end{pmatrix} , \tag{17.47}$$

where β^a are the components of β with respect to the horizontal frame basis $\{e_a\}$. The inverse of (17.47) is

$$\begin{pmatrix} e_0 \\ e_a \end{pmatrix} = \begin{pmatrix} \alpha^{-1} & -\alpha^{-1}\beta^m \\ 0 & [A^{-1}]^m_a \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial x^0} \\ \frac{\partial}{\partial x^m} \end{pmatrix} , \tag{17.48}$$

where β^m are the components of β with respect to the horizontal coordinate-induced frame basis $\{\partial/\partial x^m\}$.

The relation for the cobases dual to those in (17.47) is given by the transpose of (17.47), which we write as

$$\begin{pmatrix} \theta^0 & \theta^a \end{pmatrix} = \begin{pmatrix} \mathrm{d}x^0 & \mathrm{d}x^m \end{pmatrix} \begin{pmatrix} \alpha & \beta^a \\ 0 & A^a_m \end{pmatrix} . \tag{17.49}$$

The inverse of that is the transpose of (17.48)

$$\begin{pmatrix} \mathrm{d}x^0 & \mathrm{d}x^m \end{pmatrix} = \begin{pmatrix} \theta^0 & \theta^a \end{pmatrix} \begin{pmatrix} \alpha^{-1} & -\alpha^{-1}\beta^m \\ 0 & [A^{-1}]^m_a \end{pmatrix} . \tag{17.50}$$

Orthogonality of the e_a implies for the chart components of the spatial metric (17.37)

$$h_{mn} := h\left(\frac{\partial}{\partial x^m}, \frac{\partial}{\partial x^n}\right) = \sum_{a=1}^{3} A^a_m A^a_n \tag{17.51}$$

and its inverse

$$h^{mn} := h^{-1}(\mathrm{d}x^m, \mathrm{d}x^n) = \sum_{a=1}^{3} [A^{-1}]^m_a [A^{-1}]^n_a \,. \tag{17.52}$$

Inserting (17.49) into (17.45) and using (17.51) leads to the (3 + 1)-form of the metric in adapted coordinates

$$\begin{aligned} g = {} & (\varepsilon\alpha^2 + h(\beta, \beta))c^2 \mathrm{d}t \otimes \mathrm{d}t \\ & + c\beta_m (\mathrm{d}t \otimes \mathrm{d}x^m + \mathrm{d}x^m \otimes \mathrm{d}t) \\ & + h_{mn} \mathrm{d}x^m \otimes \mathrm{d}x^n \,, \end{aligned} \tag{17.53}$$

where $\beta_m := h_{mn}\beta^n$ are the components of $\beta^\flat := g(\beta, \cdot) = h(\beta, \cdot)$ with respect to the coordinate basis $\{\partial/\partial x^m\}$. Likewise, inserting (17.50) into (17.46) and using (17.52) leads to the (3 + 1)-form of the inverse metric in adapted coordinates (we write $\partial_t := \partial/\partial t$ and $\partial_m := \partial/\partial x^m$ for convenience)

$$\begin{aligned} g^{-1} = {} & \varepsilon c^{-2}\alpha^{-2}\partial_t \otimes \partial_t \\ & - \varepsilon c^{-1}\alpha^{-2}\beta^m (\partial_t \otimes \partial_m + \partial_m \otimes \partial_t) \\ & + (h^{mn} + \varepsilon\beta^m\beta^n)\, \partial_m \otimes \partial_n \,. \end{aligned} \tag{17.54}$$

Finally, we note that the volume form on spacetime also easily follows from (17.49)

$$\begin{aligned} \mathrm{d}\mu_g &= \theta^0 \wedge \theta^1 \wedge \theta^2 \wedge \theta^3 \\ &= \alpha\sqrt{\det\{h_{mn}\}} c\, \mathrm{d}t \wedge \mathrm{d}^3x \,, \end{aligned} \tag{17.55}$$

where we use the standard shorthand $\mathrm{d}^3x = \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3$.

17.4.2 Decomposition of the Covariant Derivative

Given horizontal vector fields X and Y, the covariant derivative of Y with respect to X need not be horizontal. Its decomposition is written as

$$\nabla_X Y = D_X Y + nK(X, Y) \,, \tag{17.56}$$

where

$$D_X Y := P^{\|}\nabla_X Y \,, \tag{17.57}$$

$$K(X, Y) := \varepsilon g(n, \nabla_X Y) \,. \tag{17.58}$$

The map D defines a covariant derivative (in the sense of Kozul; compare [17.29, Vol. 2]) for horizontal vector fields, as a trivial check of the axioms reveals. Moreover, since the commutator $[X, Y]$ of two horizontal vector fields is always horizontal (since the horizontal distribution is integrable by construction), we have

$$\begin{aligned} T^D(X, Y) &= D_X Y - D_Y X - [X, Y] \\ &= P^{\|}(\nabla_X Y - \nabla_Y X - [X, Y]) \\ &= 0 \,, \end{aligned} \tag{17.59}$$

due to ∇ being torsion free. We recall that torsion is a tensor field $T \in \Gamma T^1_2 M$ associated with each covariant derivative ∇ via

$$T^\nabla(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y] \,. \tag{17.60}$$

We have $T(X, Y) = -T(Y, X)$. As usual, even though the operations on the right-hand side of (17.60) involve tensor *fields* (we need to differentiate), the result of the operation just depends on X and Y pointwise. This one proves by simply checking the validity of $T(fX, Y) = fT(X, Y)$ for all smooth functions f. Hence, (17.59) shows that D is torsion free because ∇ is torsion free.

Finally, we can uniquely extend D to all horizontal tensor fields by requiring the Leibniz rule. Then, for X, Y, Z horizontal

$$\begin{aligned} & (D_X h)(Y, Z) \\ & = X(h(Y, Z)) - h(D_X Y, Z) - h(Y, D_X Z) \\ & = X(g(Y, Z)) - g(\nabla_X Y, Z) - g(Y, \nabla_X Z) \\ & = (\nabla_X g)(Y, Z) = 0 \,, \end{aligned} \tag{17.61}$$

due to the metricity, $\nabla g = 0$, of ∇. Hence, D is metric in the sense that

$$Dh = 0 \,. \tag{17.62}$$

The map K from pairs of horizontal vector fields (X, Y) into functions defines a symmetric tensor field. Symmetry follows from the vanishing torsion of ∇, since then

$$\begin{aligned} K(X, Y) &= \varepsilon g(n, \nabla_X Y) \\ &= \varepsilon g(n, \nabla_Y X + [X, Y]) \\ &= \varepsilon g(n, \nabla_Y X) \\ &= K(Y, X) \,, \end{aligned} \tag{17.63}$$

for horizontal X, Y. From (17.58), one sees that $K(fX, Y) = fK(X, Y)$ for any smooth function f. Hence, K defines a unique symmetric tensor field on M by stipulating that it be horizontal, i. e. $K(n, \cdot) = 0$. It is called the *extrinsic curvature* of the foliation or *second fundamental form*, the first fundamental form being the metric. From (17.58) and the symmetry just shown, one immediately infers the alternative expressions

$$K(X, Y) = -\varepsilon g(\nabla_X n, Y) = -\varepsilon g(\nabla_Y n, X) . \quad (17.64)$$

This shows the relation between the extrinsic curvature and the *Weingarten map*, **Wein**, also called the *shape operator*, which sends horizontal vectors to horizontal vectors according to

$$X \mapsto \mathbf{Wein}(X) := \nabla_X n . \quad (17.65)$$

Horizontality of $\nabla_X n$ immediately follows from n being normalized: $g(n, \nabla_X n) = \frac{1}{2} X\,(g(n, n)) = 0$. Hence, (17.64) simply becomes

$$K(X, Y) = -\varepsilon h\,(\mathbf{Wein}(X), Y) = -\varepsilon h\,(X, \mathbf{Wein}(Y)) \; , \quad (17.66)$$

where we replaced g with h – defined in (17.37) – since both entries are horizontal. It says that K is $(-\varepsilon)$ times the covariant tensor corresponding to the Weingarten map, and that the symmetry of K is equivalent to the self adjointness of the Weingarten map with respect to h. The Weingarten map characterizes the bending of the embedded hypersurface in the ambient space by answering the following question: in what direction and by what amount does the normal to the hypersurface tilt if, starting at point p, you progress within the hypersurface by the vector X. The answer is just $\mathbf{Wein}_p(X)$. Self adjointness of **Wein** then means that there always exist three ($n-1$ in general) perpendicular directions in the hypersurface along which the normal tilts in the same direction. These are the *principal curvature directions* mentioned above. The principal curvatures are the corresponding eigenvalues of **Wein**.

Finally, we note that the covariant derivative of the normal field n can be written in terms of the acceleration and the Weingarten map as follows

$$\nabla n = \varepsilon n^\flat \otimes a + \mathbf{Wein} . \quad (17.67)$$

Recalling (17.66), the purely covariant version of this is

$$\nabla n^\flat = -\varepsilon \left(K - n^\flat \otimes a^\flat \right) . \quad (17.68)$$

From (17.37) and (17.68), we derive by standard manipulation, using vanishing torsion

$$L_n h = -2\varepsilon K . \quad (17.69)$$

In the presence of torsion there would be an additional term $+2(i_n T)^\flat_s$, where the subscript s denotes symmetrization; in coordinates $[(i_n T)^\flat_s]_{\mu\nu} = n^\lambda T^\alpha_{\lambda(\mu} g_{\nu)\alpha}$.

17.5 Curvature Tensors

We wish to calculate the (intrinsic) curvature tensor of ∇ and express it in terms of the curvature tensor of D, the extrinsic curvature K, and the spatial and normal derivatives of n and K. Before we do this, we wish to say a few words on the definition of the curvature measures in general.

All notions of curvature eventually reduce to that of curves. For a surface S embedded in $\mathbb{R}^3$, we have the notion of Gauss curvature, which comes about as follows: consider a point $p \in S$ and a unit vector v at p tangent to S. Consider all smooth curves passing through p with unit tangent v. It is easy to see that the curvatures at p of all such curves are not bounded from above (due to the possibility of bending within the surface), but there will be a lower bound, $k(p, v)$, which just depends on the chosen point p and the tangent direction represented by v. Now consider $k(p, v)$ as a function of v. As v varies over all tangent directions, $k(p, v)$ will assume a minimal and a maximal value, denoted by $k_{\min}(p) = k(p, v_{\min})$ and $k_{\max}(p) = k(p, v_{\max})$, respectively. These are called the principal curvatures of S at p and their reciprocals are called the principal radii. It is clear that the principal directions $v_{\min}$ and $v_{\max}$ just span the eigenspaces of the Weingarten map discussed above. In particular, $v_{\min}$ and $v_{\max}$ are orthogonal. The Gaussian curvature $K(p)$ of S at p is then defined to be the product of the principal curvatures

$$K(p) = k_{\min}(p) \cdot k_{\max}(p) . \quad (17.70)$$

This definition is extrinsic in the sense that essential use is made of the ambient $\mathbb{R}^3$ in which S is embedded.

However, Gauss' *theorema egregium* states that this notion of curvature can also be defined intrinsically, in the sense that the value $K(p)$ can be obtained from geometric operations entirely carried out *within* the surface S. More precisely, it is a function of the first fundamental form (the metric) only, which encodes the intrinsic geometry of S, and does not involve the second fundamental form (the extrinsic curvature), which encodes how S in embedded into $\mathbb{R}^3$.

Let us briefly state Gauss' theorem in mathematical terms. Let

$$g = g_{ab}\,\mathrm{d}x^a \otimes \mathrm{d}x^b \tag{17.71}$$

be the metric of the surface in some coordinates, and

$$\Gamma^c_{ab} = \tfrac{1}{2} g^{cd}(-\partial_d g_{ab} + \partial_a g_{bd} + \partial_b g_{da}) \tag{17.72}$$

certain combinations of first derivatives of the metric coefficients, now known under the name of *Christoffel symbols*. Next we form even more complicated combinations of first and second derivatives of the metric coefficients, namely

$$R^a{}_{bcd} = \partial_c \Gamma^a_{db} - \partial_d \Gamma^a_{cb} + \Gamma^a_{cn}\Gamma^n_{db} - \Gamma^a_{dn}\Gamma^n_{cb}\,, \tag{17.73}$$

which are now known as components of the *Riemann tensor*. From them we form the totally covariant (all indices down) components

$$R_{abcd} = g_{an} R^n{}_{bcd}\,. \tag{17.74}$$

They are anti-symmetric in the first and second index pair: $R_{abcd} = -R_{bacd} = -R_{abdc}$, so that R_{1212} is the only independent component. Gauss' theorem now states that at each point on S we have

$$K = \frac{R_{1212}}{g_{11}g_{22} - g_{12}^2}\,. \tag{17.75}$$

An important part of the theorem is to show that this actually makes sense, i. e., that the right-hand side is independent of the coordinate system that we use to express the coefficients. This is easy to check once one knows that R_{abcd} are the coefficients of a tensor with the symmetries just stated. In this way the curvature of a surface, which was primarily defined in terms of curvatures of certain curves on the surface, can be understood intrinsically. In what follows we will see that the various measures of intrinsic curvatures of n-dimensional manifolds can be reduced to that of two-dimensional submanifolds, which will be called *sectional curvatures*.

Back to the general setting, we start from the notion of a covariant derivative ∇. Its associated curvature tensor is defined by

$$R(X,Y)Z = \left(\nabla_X\nabla_Y - \nabla_Y\nabla_X - \nabla_{[X,Y]}\right) Z\,. \tag{17.76}$$

For each point $p \in M$, it should be thought of as a map that assigns to each pair $X, Y \in T_pM$ of tangent vectors at p a linear map $R(X,Y) : T_pM \to T_pM$. This assignment is anti-symmetric, i. e. $R(X,Y) = -R(Y,X)$. If $R(X,Y)$ is applied to Z, the result is given by the right-hand side of (17.76). Despite first appearance, the right-hand side of (17.76) at a point $p \in M$ only depends on the values of X, Y, and Z at that point and hence defines a tensor field. This one again proves by showing the validity of $R(fX,Y)Z = R(X,fY)Z = R(X,Y)fZ = fR(X,Y)Z$ for all smooth real-valued functions f on M. In other words: all terms involving derivatives of f cancel.

From (17.76) and using (17.60), one may show that the Riemann tensor always obeys the first and second *Bianchi identities*

$$\sum_{(XYZ)} R(X,Y)Z = \sum_{(XYZ)} \{(\nabla_X T)(Y,Z) - T(X, T(Y,Z))\}\,, \tag{17.77a}$$

$$\sum_{(XYZ)} (\nabla_X R)(Y,Z) = \sum_{(XYZ)} R(X, T(Y,Z))\,, \tag{17.77b}$$

where the sums are over the three cyclic permutations of X, Y, and Z. For zero torsion these identities read in component form

$$\sum_{(\lambda\mu\nu)} R^\alpha{}_{\lambda\mu\nu} = 0\,, \tag{17.78a}$$

$$\sum_{(\lambda\mu\nu)} \nabla_\lambda R^\alpha{}_{\beta\mu\nu} = 0\,. \tag{17.78b}$$

The second traced on (α, μ) and contracted with $g^{\beta\nu}$ yields (-2) times (17.13).

The covariant Riemann tensor is defined by

$$\mathbf{Riem}(W,Z,X,Y) := g\,(W, R(X,Y)Z)\,. \tag{17.79}$$

For general covariant derivatives its only symmetry is the anti-symmetry in the last pair. But, for special choices, it acquires more. In standard general relativity we assume the covariant derivative to be *metric compatible* and *torsion free*

$$\nabla g = 0 . \quad (17.80)$$
$$T = 0 . \quad (17.81)$$

In that case the Riemann tensor has the symmetries

$$\mathbf{Riem}(W,Z,X,Y) = -\mathbf{Riem}(W,Z,Y,X) , \quad (17.82a)$$
$$\mathbf{Riem}(W,Z,X,Y) = -\mathbf{Riem}(Z,W,X,Y) , \quad (17.82b)$$
$$\mathbf{Riem}(W,X,Y,Z) + \mathbf{Riem}(W,Y,Z,X) + \mathbf{Riem}(W,Z,Y,X) = 0 , \quad (17.82c)$$
$$\mathbf{Riem}(W,Z,X,Y) = R(X,Y,W,Z) . \quad (17.82d)$$

Equation (17.82a) is true by definition (17.76), (17.82b) is equivalent to metricity of ∇, and (17.82c) is the first Bianchi identity in the case of zero torsion. The last symmetry (17.82d) is a consequence of the preceding three. Together (17.82a), (17.82b), and (17.82d) say that, at each point $p \in M$, **Riem** can be thought of as a symmetric bilinear form on the anti-symmetric tensor product $T_pM \wedge T_pM$. The latter has dimension $N = \frac{1}{2}n(n-1)$ if M has dimension n, and the space of symmetric bilinear forms has dimension $\frac{1}{2}N(N+1)$. From that number we have to subtract the number of independent conditions (17.82c), which is $\binom{n}{4}$ in dimensions $n \geq 4$ and zero otherwise. Indeed, it is easy to see that (17.82c) is identically satisfied as a consequence of (17.82a) and (17.82b) if any two vectors W, Z, X, Y coincide (proportionality is sufficient). Hence, the number # of independent components of the curvature tensor is

$$\#\mathbf{Riem} = \begin{cases} \frac{1}{2}N(N+1) - \binom{n}{4} = \frac{1}{12}n^2(n^2-1) & \text{for} \quad n \geq 4 , \\ 6 & \text{for} \quad n = 3 , \\ 1 & \text{for} \quad n = 2; . \end{cases} \quad (17.83)$$

The Ricci and scalar curvatures are obtained by taking traces with respect to g: let $\{e_1, \dots, e_n\}$ be an orthonormal basis and $g(e_a, e_b) = \delta_{ab}\varepsilon_a$ (no summation) with $\varepsilon_a = \pm 1$; then

$$\mathbf{Ric}(X,Y) = \sum_{a=1}^{n} \varepsilon_a \, \mathbf{Riem}(e_a, X, e_a, Y) , \quad (17.84)$$
$$\mathbf{Scal} = \sum_{a=1}^{n} \varepsilon_a \, \mathbf{Ric}(e_a, e_a) . \quad (17.85)$$

The Einstein tensor is

$$\mathbf{Ein} = \mathbf{Ric} - \tfrac{1}{2}\,\mathbf{Scal}\, g . \quad (17.86)$$

The sectional curvature is defined by

$$\mathbf{Sec}(X,Y) = \frac{\mathbf{Riem}(X,Y,X,Y)}{g(X,X)g(Y,Y) - [g(X,Y)]^2} . \quad (17.87)$$

Here X, Y is a pair of linearly independent tangent vectors that span a two-dimensional tangent subspace restricted to which g is nondegenerate. We will say that span$\{X, Y\}$ is nondegenerate. This is the necessary and sufficient condition for the denominator on the right-hand side to be nonzero. The quantity $\mathbf{Sec}(X, Y)$ is called the *sectional curvature* of the manifold (M, g) at point p tangent to span$\{X, Y\}$. From the symmetries of **Riem**, it is easy to see that the right-hand side of (17.87) does indeed only depend on the span of X, Y. That is, for any other pair X', Y' such that span$\{X', Y'\}$ = span$\{X, Y\}$, we have $\mathbf{Sec}(X', Y') = \mathbf{Sec}(X, Y)$. The geometric interpretation of $\mathbf{Sec}(X, Y)$ is as follows: consider all geodesics of (M, g) that pass through the considered point $p \in M$ in a direction tangential to span$\{X, Y\}$. In a neighborhood of p they form an embedded 2-surface in M whose Gauss curvature is just $\mathbf{Sec}(X, Y)$.

Now, **Riem** is determined by components of the form $\mathbf{Riem}(X, Y, X, Y)$, as follows from the fact that **Riem** is a *symmetric* bilinear form on $TM \wedge TM$. This remains true if we restrict to those X, Y whose span is nondegenerate, since they lie dense in $TM \wedge TM$ and $(X, Y) \mapsto \mathbf{Riem}(X, Y, X, Y)$ is continuous. This shows that the full information of the Riemann tensor can be reduced to certain Gauss curvatures.

This also provides a simple geometric interpretation of the scalar and Einstein curvatures in terms of sectional curvatures. Let $\{X_1, \dots, X_n\}$ be any set of pairwise-orthogonal nonnull vectors. The $\frac{1}{2}n(n-1)$ planes span$\{X_a, X_b\}$ are nondegenerate and also pairwise orthogonal. It then follows from (17.85) and (17.87) that the scalar curvature is twice the sum of all

sectional curvatures

$$\mathbf{Scal} = 2 \sum_{\substack{a,b=1 \\ a<b}}^{n} \mathbf{Sec}(X_a, X_b) \ . \tag{17.88}$$

The sum on the right-hand side of (17.88) is the same for any set of $\frac{1}{2}n(n-1)$ nondegenerate and pairwise-orthogonal 2-planes. Hence, the scalar curvature can be said to be twice the sum of scalar curvatures, or $n(n-1)$ times the mean scalar curvature. Similarly for the Ricci and Einstein curvatures. The symmetry of the Ricci and Einstein tensors implies that they are fully determined by their components $\mathbf{Ric}(W, W)$ and $\mathbf{Ein}(W, W)$. Again, this remains true if we restrict to the dense set of nonnull W, i. e. $g(W, W) \neq 0$. Let now $\{X_1, \ldots, X_{n-1}\}$ be any set of mutually orthogonal vectors (again they need not be normalized) in the orthogonal complement of W. As before, the $\frac{1}{2}(n-1)(n-2)$ planes span$\{X_a, X_b\}$ are nondegenerate and pairwise orthogonal. From (17.84), (17.86), and (17.87), it follows that

$$\mathbf{Ric}(W, W) = g(W, W) \sum_{a=1}^{n-1} \mathbf{Sec}(W, X_a) \tag{17.89}$$

and

$$\mathbf{Ein}(W, W) = -g(W, W) \sum_{\substack{a,b=1 \\ a<b}}^{n-1} \mathbf{Sec}(X_a, X_b) \ . \tag{17.90}$$

Again, the right-hand sides will be the same for any set $\{X_1, \ldots, X_{n-1}\}$ of $n-1$ mutually orthogonal vectors in the orthogonal complement of W. Note that $\mathbf{Ric}(W, W)$ involves all sectional curvatures involving W, whereas $\mathbf{Ein}(W, W)$ involves all sectional curvatures orthogonal to W. For normalized W, where $g(W, W) = \sigma = \pm 1$, we can say that $-\sigma G(W, W)$ is the sum of sectional curvatures orthogonal to W, or $\frac{1}{2}(n-1)(n-2)$ times their mean. Note that for timelike W we have $\sigma = -1$ and $G(W, W)$ is just the sum of spatial sectional curvatures.

Finally, we mention the Weyl tensor, which contains that part of the information in the curvature tensor not captured by the Ricci (or Einstein) tensor. To state its form in a compact form, we introduce the *Kulkarni–Nomizu product*, denoted by an encircled wedge, $\owedge$, which is a bilinear symmetric product on the space of covariant symmetric rank-two tensors with values in the covariant rank-four tensors that have the symmetries (17.82) of the Riemann tensor. Let k and ℓ be two symmetric covariant second-rank tensors; then their Kulkarni–Nomizu product is defined by

$$\begin{aligned} k \owedge \ell(X_1, X_2, X_3, X_4) := {} & k(X_1, X_3)\ell(X_2, X_4) \\ & + k(X_2, X_4)\ell(X_1, X_3) \\ & - k(X_1, X_4)\ell(X_2, X_3) \\ & - k(X_2, X_3)\ell(X_1, X_4) \ , \end{aligned} \tag{17.91}$$

or in components

$$(k \owedge \ell)_{abcd} = k_{ac}\ell_{bd} + k_{bd}\ell_{ac} - k_{ad}\ell_{bc} - k_{bc}\ell_{ad} \ . \tag{17.92}$$

The Weyl tensor **Weyl** is of the same type as **Riem** but in addition totally trace free. Its definition is

$$\mathbf{Weyl} = \mathbf{Riem} - \frac{1}{n-2}\left(\mathbf{Ric} - \frac{1}{2(n-1)}\,\mathbf{Scal}\, g\right) \owedge g \ . \tag{17.93}$$

Its number of independent components is

$$\#\mathbf{Weyl} = \begin{cases} \frac{1}{12}n(n+1)\,[n(n-1)-6] & \text{for} \quad n \geq 4 \ , \\ 0 & \text{for} \quad n \leq 3 \ . \end{cases} \tag{17.94}$$

Note that in $n = 3$ dimensions the Weyl tensor also always vanishes, so that (17.93) can be used to express the Riemann tensor in terms of the Ricci and scalar curvature

$$\mathbf{Riem} = \left(\mathbf{Ric} - \tfrac{1}{4}\mathbf{Scal}\, g\right) \owedge g \quad (\text{for } n = 3) \ . \tag{17.95}$$

A metric manifold (M, g) is said to be of *constant curvature* if

$$\mathbf{Riem} = k g \owedge g \ , \tag{17.96}$$

for some function k. Then $\mathbf{Ric} = 2k(n-1)g$ and $\mathbf{Ein} = -k(n-1)(n-2)g$. We recall that manifolds (M, g) for which the Einstein tensor (equivalently, the Ricci tensor) is pointwise proportional to the metric are called *Einstein spaces*. The twice-contracted second Bianchi identity (17.13) shows that k must be a constant unless $n = 2$. For $n = 3$, (17.95) shows that Einstein spaces are of constant curvature.

17.5.1 Comparing Curvature Tensors

Sometimes one wants to compare two different curvature tensors belonging to two different covariant derivatives $\hat{\nabla}$ and ∇. In what follows, all quantities referring to $\hat{\nabla}$ carry a hat. Recall that a covariant derivative can be considered as a map $\nabla : \Gamma T^1_0 M \times \Gamma T^1_0 M \to \Gamma T^1_0 M$, $(X, Y) \mapsto \nabla_X Y$, which is $C^\infty(M)$-linear in the first and a derivation in the second argument. That is, for $f \in C^\infty(M)$ we have $\nabla_{fX+Y} Z = f\nabla_X Z + \nabla_Y Z$ and $\nabla_X(fY + Z) = X(f)Y + f\nabla_X Y + \nabla_X Z$. This implies that the difference of two covariant derivatives is $C^\infty(M)$-linear also in the second argument and hence a tensor field

$$\hat{\nabla} - \nabla =: \Delta \in \Gamma T^1_2 M \, . \tag{17.97}$$

Replacing $\hat{\nabla}$ with $\nabla + \Delta$ in the definition of the curvature tensor for $\hat{\nabla}$ according to (17.76) directly leads to

$$\begin{aligned}\hat{R}(X, Y)Z = R(X, Y)Z & \\ &+ (\nabla_X \Delta)(Y, Z) - (\nabla_Y \Delta)(X, Z) \\ &+ \Delta\,(X, \Delta(Y, Z)) - \Delta\,(Y, \Delta(X, Z)) \\ &+ \Delta\,(T(X, Y), Z) \; .\end{aligned} \tag{17.98}$$

Note that so far no assumptions have been made concerning torsion or metricity of $\hat{\nabla}$ and ∇. This formula is generally valid. In the special case where $\hat{\nabla}$ and ∇ are the Levi-Civita covariant derivatives with respect to two metrics $\hat{g}$ and g, we set (for the rest of this subsection, h has a different meaning than that of (17.37))

$$h := \hat{g} - g \, , \tag{17.99}$$

which is a symmetric covariant tensor field. We recall that the Levi-Civita covariant derivative is uniquely determined by the metric. For ∇, this reads

$$\begin{aligned}&2g(\nabla_X Y, Z) \\ &= X(g(Y, Z)) + Y(g(Z, X)) - Z(g(X, Y)) \\ &\quad - g(X, [Y, Z]) + g(Y, [Z, X]) + g(Z, [X, Y]) \, .\end{aligned} \tag{17.100}$$

Subtracting (17.100) from the corresponding formula with ∇ and g replaced by $\hat{\nabla}$ and $\hat{g}$ yields, using $T = 0$,

$$\begin{aligned}&2\hat{g}\,(\Delta(X, Y), Z) \\ &= -(\nabla_Z h)(X, Y) + (\nabla_X h)(Y, Z) + (\nabla_Y h)(Z, X) \, .\end{aligned} \tag{17.101}$$

This formula expresses Δ as a functional of g and $\hat{g}$. There are various equivalent forms of it. We have chosen a representation that somehow minimizes the appearance of $\hat{g}$. Note that g enters in h as well as ∇, whereas $\hat{g}$ enters in h and via the scalar product on the left-hand side. The latter obstructs expressing Δ as a functional of g and h alone. In components (17.101) reads

$$\Delta^a_{bc} = \tfrac{1}{2}\hat{g}^{an}\left(-\nabla_n h_{bc} + \nabla_b h_{cn} + \nabla_c h_{nb}\right) \, . \tag{17.102}$$

Note that one could replace the components of h with those of $\hat{g} = g + h$ in the bracket on the right-hand side, since the covariant derivatives of g vanish.

Now suppose we consider h and its first and second derivatives to be small and we wanted to know the difference in the covariant derivatives and curvatures only to leading (linear) order in h. To that order, we may replace $\hat{g}$ with g on the left-hand side of (17.101) and the right-hand side of (17.102). Moreover, we may neglect the Δ-squared terms in (17.98) and obtain, writing δR for the first order in h contribution to $\hat{R} - R$

$$\delta R^a{}_{bcd} = \nabla_c \Delta^a_{db} - \nabla_d \Delta^a_{cb} \, . \tag{17.103}$$

From this, the first-order variation of the Ricci tensor follows

$$\begin{aligned}\delta R_{ab} &= \nabla_n \Delta^n_{ab} - \nabla_b \Delta^n_{na} \\ &= \tfrac{1}{2}\left(-\Delta_g h_{ab} - \nabla_a \nabla_b h + \nabla_a \nabla^n h_{nb} + \nabla_b \nabla^n h_{na}\right) ,\end{aligned} \tag{17.104}$$

where $\Delta_g := g^{ab}\nabla_a\nabla_b$ and $h = g^{ab}h_{ab}$. Finally, the variation of the scalar curvature is (note that $\delta g^{ab} = -g^{ac}g^{bd}\delta g_{cd} = -h^{ab}$)

$$\delta R = -R_{ab}h^{ab} + \nabla_a U^a \, , \tag{17.105a}$$

where

$$\begin{aligned}U^a &= g^{nm}\Delta^a_{nm} - g^{an}\Delta^m_{mn} \\ &= \nabla_b h^{ab} - \nabla^a h^b_b \\ &= G^{abcd}\nabla_b h_{cd} \, .\end{aligned} \tag{17.105b}$$

Here we made use of the *DeWitt metric*, which defines a symmetric nondegenerate bilinear form on the space of symmetric covariant rank-two tensors and which in components reads

$$G^{abcd} = \frac{1}{2}\left(g^{ac}g^{bd} + g^{ad}g^{bc} - 2g^{ab}g^{cd}\right) . \tag{17.106}$$

We will later have to say more about it.

We also wish to state a useful formula that compares the curvature tensors for conformally related metrics, i.e.

$$\hat{g} = e^{2\phi} g\,, \tag{17.107}$$

where $\phi : M \to \mathbb{R}$ is smooth. Then

$$\begin{aligned}\mathbf{Riem}_{\hat{g}} = e^{2\phi}\big[&\mathbf{Riem}_g \\ &- (\nabla\nabla\phi) \owedge g + (d\phi \otimes d\phi) \owedge g \\ &- \tfrac{1}{2} g^{-1}(d\phi, d\phi) g \owedge g\big]\,.\end{aligned} \tag{17.108}$$

(This is best proven by using Cartan's structure equation.) From that equation, we deduce the transformation properties of the Ricci tensor

$$\begin{aligned}\mathbf{Ric}_{\hat{g}} = \mathbf{Ric}_g &- \big(\Delta_g\phi + (n-2) g^{-1}(d\phi, d\phi)\big) g \\ &- (n-2)(\nabla\nabla\phi - d\phi \otimes d\phi)\,,\end{aligned} \tag{17.109}$$

where, as above, Δ_g is the Laplacian/d'Alembertian for g. For the scalar curvature, we get

$$\begin{aligned}\mathbf{Scal}_{\hat{g}} = e^{-2\phi}\big(&\mathbf{Scal}_g - 2(n-1)\Delta_g\phi \\ &- (n-1)(n-2) g^{-1}(d\phi, d\phi)\big)\,.\end{aligned} \tag{17.110}$$

This law has a linear dependence on the second and a quadratic dependence on the first derivatives of ϕ. If the conformal factor is written as an appropriate power of some positive function $\Omega : M \to \mathbb{R}_+$, we can eliminate all dependence on first and just retain the second derivatives. In $n > 2$ dimensions it is easy to check that the rule is this

$$e^{2\phi} = \Omega^{\frac{4}{n-2}}\,; \tag{17.111}$$

then (17.110) becomes

$$\mathbf{Scal}_{\hat{g}} = -\frac{4(n-1)}{n-2}\Omega^{-\frac{n+2}{n-2}} \mathcal{D}_g \Omega\,, \tag{17.112a}$$

where

$$\mathcal{D}_g = \Delta_g - \frac{n-2}{4(n-1)}\mathbf{Scal}_g\,. \tag{17.112b}$$

Here $\mathcal{D}_g$ is a linear differential operator which is elliptic for Riemannian and hyperbolic for Lorentzian metrics g. If we set $\Omega = \Omega_1\Omega_2$ and apply (17.112) twice, one time to the pair $(\hat{g}, g)$ and the other time to $(\hat{g}, \Omega_2 g)$, we obtain by direct comparison (and renaming Ω_2 to Ω thereafter) the conformal transformation property for the operator $\mathcal{D}_g$

$$\mathcal{D}_{\Omega^{\frac{4}{n-2}} g} = M\left(\Omega^{-\frac{n+2}{n-2}}\right) \circ \mathcal{D}_g \circ M(\Omega)\,, \tag{17.113}$$

where $M(\Omega)$ is the linear operator of multiplication with Ω. This is the reason why $\mathcal{D}_g$ is called the *conformally covariant Laplacian* (for Riemannian g) or the *conformally covariant wave operator* (for Lorentzian g). As we will see, it has useful applications to the initial-data problem in GR.

17.5.2 Curvature Decomposition

Using (17.56), we can decompose the various curvature tensors. From now on, h will again denote the horizontal projection (17.37) of the metric g. First we let X, Y, Z be horizontal vector fields. We use (17.56) in (17.76) and get the general formula (i. e. not yet making use of the fact that ∇ and D are metric and torsion free)

$$\begin{aligned}R(X,Y)Z = R^D(X,Y)Z &+ (\nabla_X n)K(Y,Z) - (\nabla_Y n)K(X,Z) \\ &+ n[(D_X K)(Y,Z) - (D_Y K)(X,Z)] \\ &+ nK(T^D(X,Y),Z)\,,\end{aligned} \tag{17.114}$$

where

$$R^D(X,Y)Z := \big(D_X D_Y - D_Y D_X - D_{[X,Y]}\big) Z \tag{17.115}$$

is the horizontal curvature tensor associated with the Levi-Civita covariant derivative D of h. This formula is general in the sense that it is valid for any covariant derivative. No assumptions have been made so far concerning metricity or torsion, and this is why the torsion T^D of D (defined in (17.59)) makes an explicit appearance. From now on we shall restrict to vanishing torsion. We observe that the first two lines on the right-hand side of (17.114) are horizontal, whereas the last two lines are proportional to n. Decomposition into horizontal and normal components, respectively, leads

to ($T^D = 0$)

$$\begin{aligned}&\mathbf{Riem}(W,Z,X,Y)\\&= \mathbf{Riem}^D(W,Z,X,Y)\\&\quad -\varepsilon[K(W,X)K(Z,Y) - K(W,Y)K(Z,X)]\,,\end{aligned} \tag{17.116}$$

where we used $h(W,\nabla_X n) = -\varepsilon K(W,X)$ from (17.64), and

$$\begin{aligned}&\mathbf{Riem}(n,Z,X,Y)\\&= \varepsilon\,[(D_X K)(Y,Z) - (D_Y K)(X,Z)]\;.\end{aligned} \tag{17.117}$$

The remaining curvature components are those involving two entries in the n direction. Using (17.68), we obtain via standard manipulations (now using metricity and vanishing torsion)

$$\begin{aligned}\mathbf{Riem}(X,n,Y,n) &= i_X\left(\nabla_Y\nabla_n - \nabla_n\nabla_Y - \nabla_{[Y,n]}\right)n^\flat\\&= i_X i_Y\left(\varepsilon L_n K + K\circ K + Da^\flat - \varepsilon a^\flat\otimes a^\flat\right)\;.\end{aligned} \tag{17.118}$$

Here $K\circ K(X,Y) := h^{-1}(i_X K, i_Y K) = i_X K((i_Y K)^\sharp)$ and we used the following relation between covariant and Lie derivatives (which will have additional terms in the case of nonvanishing torsion)

$$\nabla_n K = L_n K + 2\varepsilon K\circ K\;. \tag{17.119}$$

Note also that the left-hand side of (17.117) is symmetric as a consequence of (17.82d). On the right-hand side only $Da^\flat$ is not immediately seen to be symmetric, but that follows from (17.41b).

Equations (17.114), (17.116), and (17.117) express all components of the spacetime curvature in terms of horizontal quantities and their Lie derivatives L_n in the normal direction. According to (17.44), the latter can be replaced by a combination of Lie derivatives along the time vector field $\partial/\partial t$ and the shift β. From (17.42b), we infer that $L_{\alpha n} = \alpha L_n$ on horizontal covariant tensor fields; therefore, we may replace

$$L_n \to \alpha^{-1}\left(L_{\frac{\partial}{\partial t}} - L_\beta\right) \to \alpha^{-1}\left(L^{\|}_{\frac{\partial}{\partial t}} - L^{\|}_\beta\right) \tag{17.120}$$

on horizontal covariant tensor fields. Here we set $L^{\|} = P^{\|}\circ L$, i. e. Lie derivative (as operation in the ambient spacetime) followed by horizontal projection. Moreover, using (17.39), the acceleration 1-form $a^\flat$ may be replaced by the spatial derivative of the lapse function

$$a^\flat = -\varepsilon\alpha^{-1}D\alpha\;. \tag{17.121}$$

Hence, the combination of accelerations appearing in (17.117) may be written as

$$Da^\flat - \varepsilon a^\flat\otimes a^\flat = -\varepsilon\alpha^{-1}D^2\alpha\;. \tag{17.122}$$

Note that $D^2\alpha := DD\alpha$ is just the horizontal Hessian of α with respect to h.

17.6 Decomposing Einstein's Equations

The curvature decomposition of the previous section can now be used to decompose Einstein's equations. For this we decompose the Einstein tensor **Ein** into the normal–normal, normal–tangential, and tangential–tangential parts. Let $\{e_0, e_1, e_2, e_3\}$ be an orthonormal frame with $e_0 = n$, i. e. adapted to the foliation as in Sect. 17.4.1. Then (17.90) together with (17.116) immediately lead to

$$2\mathbf{Ein}(e_0,e_0) = -\left[K_{ab}K^{ab} - \left(K^a_a\right)^2\right] - \varepsilon\mathbf{Scal}^D\,, \tag{17.123}$$

where $\mathbf{Scal}^D$ is the scalar curvature of D, i. e. of the spacelike leaves in the metric h. Similarly, we obtain from (17.117)

$$\mathbf{Ein}(e_0,e_a) = \mathbf{Ric}(e_0,e_a) = -\varepsilon\left[D^b K_{ab} - D_a K^b_b\right]\;. \tag{17.124}$$

The normal–normal component of the Ricci tensor cannot likewise be expressed simply in terms of horizontal quantities, the geometric reason being that, unlike the Einstein tensor, it involves nonhorizontal sectional curvatures (compare (17.89) and (17.90)).

A useful expression follows from taking the trace of (17.118), considered as a symmetric bilinear form in X and Y. The result is

$$\mathbf{Ric}(e_0, e_0) = -K_{ab}K^{ab} + \left(K^c_c\right)^2 + \varepsilon\nabla\cdot V\ , \tag{17.125}$$

where $\nabla\cdot$ denotes the divergence with respect to ∇ and V is a vector field on M whose normal component is the trace of the extrinsic curvature and whose horizontal component is ε times the acceleration on n

$$V = nK^c_c + \varepsilon a\ . \tag{17.126}$$

For the horizontal–horizontal components of Einstein's equations, it turns out to be simpler to use their alternative form (17.6b) with the Ricci tensor on the left-hand side. For that, we need the horizontal components of the Ricci tensor, which we easily get from (17.116) and (17.118)

$$\begin{aligned}\mathbf{Ric}(e_a, e_b) = {}&\mathbf{Ric}^D(e_a, e_b)\\ &+ L_nK_{ab} + 2\varepsilon K_{ac}K^c_b - \varepsilon K_{ab}K^c_c\\ &+ \varepsilon D_a a_b - a_a a_b\ .\end{aligned} \tag{17.127}$$

For later applications we also note the expression for the scalar curvature. It follows, e.g., from adding the horizontal trace of (17.127) to ε times (17.125). This leads to

$$\mathbf{Scal} = \mathbf{Scal}^D - \varepsilon\left[K_{ab}K^{ab} - \left(K^a_a\right)^2\right] + 2\nabla\cdot V\ . \tag{17.128}$$

Here we made use of the relation between the ∇ and D derivatives for the acceleration 1-form

$$\nabla a^\flat = Da^\flat + \varepsilon n^\flat \otimes \nabla_n a^\flat + i_n K \otimes n^\flat\ , \tag{17.129}$$

whose trace gives the following relation between the ∇ and D divergences of a

$$\nabla\cdot a = D\cdot a - \varepsilon h(a, a)\ . \tag{17.130}$$

Another possibility would have been to use (17.123) and (17.125) in $\mathbf{Scal} = -2\varepsilon(\mathbf{Ein}(e_0, e_0) - \mathbf{Ric}(e_0, e_0))$.

Using (17.123) and (17.124), and also using the DeWitt metric (17.106) for notational ease, we can immediately write down the normal–normal and normal–tangential components of Einstein's equations (17.3)

$$G^{abcd}K_{ab}K_{cd} + \varepsilon\mathbf{Scal}^D = -2\kappa\mathbf{T}(n, n)\ , \tag{17.131a}$$

$$G^{abcd}D_bK_{cd} = -\varepsilon\kappa h^{ab}\mathbf{T}(n, e_b)\ . \tag{17.131b}$$

From (17.66) and (17.106), we notice that the bilinear form on the left-hand side of (17.131a) can be written as

$$\begin{aligned}G(K, K) :&= G^{abcd}K_{ab}K_{cd}\\ &= \mathrm{Tr}(\mathbf{Wein}\circ\mathbf{Wein}) - (\mathrm{Tr}(\mathbf{Wein}))^2\ .\end{aligned} \tag{17.132}$$

Here the trace is natural (needs no metric for its definition), since **Wein** is an endomorphism. In a local frame in which **Wein** is diagonal with entries $\vec{k} := (k_1, k_2, k_3)$, we have

$$G(K, K) := (\delta^{ab} - 3n^a n^b)k_a k_b\ , \tag{17.133}$$

where n^a are the components of the normalized vector $(1, 1, 1)/\sqrt{3}$ in eigenvalue space, which we identify with $\mathbb{R}^3$ endowed with the standard Euclidean inner product. Hence, denoting by θ the angle between $\vec{n}$ and $\vec{k}$, we have

$$G(K, K) = \begin{cases} 0 & \text{if}\quad |\cos\theta| = \dfrac{1}{\sqrt{3}}\ ,\\ >0 & \text{if}\quad |\cos\theta| < \dfrac{1}{\sqrt{3}}\ ,\\ <0 & \text{if}\quad |\cos\theta| > \dfrac{1}{\sqrt{3}}\ .\end{cases} \tag{17.134}$$

Note that $|\cos\theta| = 1/\sqrt{3}$ describes a double cone around the symmetry axis generated by $\vec{n}$ and vertex at the origin, whose opening angle is just right so as to contain all three axes of $\mathbb{R}^3$. For eigenvalue vectors inside this cone the bilinear form is negative, outside this cone positive. Positive $G(K, K)$ require sufficiently anisotropic Weingarten maps, or, in other words, sufficiently large deviations from being umbilical points.

The tangential–tangential component of Einstein's equations in the form (17.5) immediately follows from (17.127). In the ensuing formula we use (17.120) to explicitly solve for the horizontal Lie derivative of K with

respect to $\partial/c\partial t$ and also (17.122) to simplify the last two terms in (17.127). This results in

$$\begin{aligned}\dot{K}_{ab} &:= \left(L^{\|}_{\frac{\partial}{c\partial t}} K\right)_{ab} \\ &= \left(L^{\|}_{\beta} K\right)_{ab} + D_a D_b \alpha \\ &\quad + \alpha\left[-2\varepsilon K_{ac}K^c_b + \varepsilon K_{ab}K^c_c - \mathbf{Ric}^D(e_a, e_b)\right] \\ &\quad - \alpha\varepsilon\frac{\kappa}{n-2} h_{ab}\mathbf{T}(n,n) \\ &\quad + \alpha\kappa\left(\mathbf{T} - \frac{1}{n-2}\mathrm{Tr}_h(\mathbf{T})h\right)(e_a, e_b)\,. \end{aligned} \tag{17.135}$$

Note that in the last term the trace of $\mathbf{T}$ is taken with respect to h and not g. The relation is $\mathrm{Tr}_h(\mathbf{T}) = \mathrm{Tr}_g(\mathbf{T}) - \varepsilon\mathbf{T}(n,n)$.

The only remaining equation that needs to be added here is that which relates the time derivative of h with K. This we get from (17.69) and (17.120)

$$\dot{h}_{ab} := \left(L^{\|}_{\frac{\partial}{c\partial t}} h\right)_{ab} = \left(L^{\|}_{\beta} h\right)_{ab} - 2\varepsilon K_{ab}\,. \tag{17.136}$$

Equations (17.136) and (17.135) are six first-order-in-time evolution equations for the pair (h, K). This pair cannot be freely specified but has to obey the four equations (17.131a) and (17.131b) which do not contain any time derivatives of h or K. Equations (17.131a) and (17.131b) are therefore referred to as *constraints*, more specifically (17.131a) as a *scalar constraint* (also *Hamiltonian constraint*) and (17.131b) as a *vector constraint* (also *diffeomorphism constraint*).

We derived these equations from the 3 + 1 split of a spacetime that we considered to be given. Despite having expressed all equations in terms of horizontal quantities, there is still a relic of the ambient space in our equations, namely the Lie derivative with respect to $\partial/\partial ct$. We now erase this last relic by interpreting this Lie derivative as an ordinary partial derivative of some t-dependent tensor field on a genuine three-dimensional manifold Σ, which is not thought of as being embedded into a spacetime. The horizontal projection $L^{\|}_{\beta}$ of the spacetime Lie derivative that appears on the right-hand sides of the evolution equations above then translates to the ordinary intrinsic Lie derivative on Σ with respect to β. This is how from now on we shall read the above equations. Spacetime does not yet exist. Rather, it has to be constructed from the evolution of the fields according to the equations, usually complemented by the equations that govern the evolution of the matter fields. In these evolution equations α and β are freely specifiable functions, the choice of which is subject to mathematical/computational convenience. Once α and β are specified and h as a function of parameter time has been determined, we can form the expression (17.53) for the spacetime metric and know that, by construction, it will satisfy Einstein's equations.

To sum up, the initial-value problem consists in the following steps:

1. Choose a 3-manifold Σ.
2. Choose a time-parameter-dependent lapse function α and a time-parameter-dependent shift vector field β.
3. Find a Riemannian metric $h \in \Gamma T^0_2\Sigma$ and a symmetric covariant rank-two tensor field $h \in \Gamma T^0_2\Sigma$ that satisfy (17.131a) and (17.131b) either in vacuum ($\mathbf{T} = \mathbf{T}_\Lambda$; cf. (17.4)) or after specifying some matter model.
4. Evolve these data via (17.136) and (17.135), possibly complemented by the evolution equations for the matter variables.
5. Construct from the solution the spacetime metric g via (17.53).

For this to be consistent, we need to check that the evolution according to (17.136) and (17.135) will preserve the constraints (17.131a) and (17.131b). At this stage this could be checked directly, at least in the vacuum case. The easiest way to do this is to use the equivalence of these equations with Einstein's equations and then employ the twice-contracted second Bianchi identity (17.13). It follows that $\nabla_\mu E^{\mu\nu} \equiv 0$, where $E^{\mu\nu} = G^{\mu\nu} + \lambda g^{\mu\nu}$. The four constraints (17.131a) and (17.131b) are equivalent to $E^{00} = 0$ and $E^{0m} = 0$, and the six second-order equations $E^{mn} = 0$ to the 12 first-order evolution equations (17.136) and (17.135). In coordinates, the identity $\nabla_\mu E^{\mu\nu} = 0$ reads

$$\partial_0 E^{0\nu} = -\partial_m E^{m\nu} - \Gamma^{\mu}_{\mu\lambda}E^{\lambda\nu} - \Gamma^{\nu}_{\mu\lambda}E^{\mu\lambda}\,, \tag{17.137}$$

which shows immediately that the time derivatives of the constraint functions are zero if the constraints vanished initially. This suffices for analytic data, but in the general case one has to do more work. Fortunately the equations for the evolution of the constraint functions can be put into a symmetric hyperbolic form, which suffices to conclude the desired result in the general case. For a mathematically more thorough discussion of the Cauchy problem, we refer the reader to Chap. 16.

Finally, we wish to substantiate our earlier claim that any Σ can carry some initial data. Let us show this for closed Σ. To this end, we choose a matter model such that the right-hand side of (17.131b) vanishes. Note that this still allows for arbitrary cosmological constants, since $\mathbf{T}_\Lambda(n, e_a) \propto g(n, e_a) = 0$. Next we restrict to those pairs (h, K) where $K = \lambda h$ for some constant λ. Geometrically this means that, in the spacetime to be developed, the Cauchy surface will be totally umbilical (isotropic Weingarten map). Due to this proportionality and the previous assumption, the vector constraint (17.131b) will be satisfied. In the scalar constraint we have $G(K, K) = G(\lambda h, \lambda h) = -6\lambda^6$, so that it will be satisfied provided that

$$-\varepsilon \mathbf{Scal}^D = 2\kappa \mathbf{T}(n, n) - 6\lambda^2 \,. \tag{17.138}$$

For the argument to follow, the Lorentzian signature, $\varepsilon = -1$, will matter. For physical reasons we assume the weak energy condition so that $\kappa\mathbf{T}(n, n) \geq 0$, which makes a positive contribution to the right-hand side of (17.138). However, if we choose the modulus of λ sufficiently large we can make the right-hand side negative somewhere (or everywhere, since Σ is compact). Now, in dimensions three or higher the following is true [17.30, Theorem 1.1]: any smooth function on a compact manifold which is negative somewhere is the scalar curvature for some smooth Riemannian metric. Hence, a smooth h exists which solves (17.138) for any given $\mathbf{T}(n, n) \geq 0$, provided we choose $\lambda^2 > |\lambda|$ sufficiently large. If Σ is not closed, a corresponding theorem may also be shown [17.31].

The above argument crucially depends on the signs. There is no corresponding statement for positive scalar curvature. In fact, there is a strong topological obstruction against Riemannian metrics of strictly positive scalar curvature. It follows from [17.32, Theorem 8.1] that a three-dimensional closed orientable Σ allows for Riemannian metrics with positive scalar curvature iff its prime decomposition consists of prime manifolds with finite fundamental group or *handles* $S^1 \times S^2$. All manifolds whose prime list contains at least one so-called $K(\pi, 1)$-factor (a 3-manifold whose only nontrivial homotopy group is the first) are excluded. See, e.g., [17.33] for more explanation of these notions. We conclude that the given argument crucially depends on $\varepsilon = -1$.

17.6.1 A Note on Slicing Conditions

The freedom in choosing the lapse and shift functions can be of much importance, theoretically and in numerical evolution schemes. This is particularly true for the lapse function α, which determines the amount of proper length by which the Cauchy slice advances in the normal direction per unit parameter interval. If a singularity is to form in spacetime due to the collapse of matter within a bounded spatial region, it would clearly be advantageous to not let the slices run into the singularity before the outer parts of it have had any chance to develop a sufficiently large portion of spacetime that one might be interested in, e.g. for the study of gravitational waves produced in the past. This means that one would like to slow down α in regions which are likely to develop a singularity and speed up α in those regions where it seems affordable. Take as an example the *equal-speed* gauge $\alpha = 1$ and $\beta = 0$, so that $g = -c^2\,dt^2 + h$. This means that $n = \partial/\partial ct$ is geodesic. Taking such a gauge from the $t = 0$ slice in the Schwarzschild/Kruskal spacetime would let the slices run into the singularity after a proper time of $t = \pi GM/c^3$, where M is the mass of the black hole. In that short period of time the slices had no chance to explore a significant portion of spacetime outside the black hole.

A gauge condition that one may anticipate to have singularity-avoiding character is that where α is chosen such that the divergence of the normal field n is zero. This condition just means that the locally comoving infinitesimal volume elements do not change volume, for $L_n\,d\mu = (\nabla\cdot n)\,d\mu$, where $d\mu = \det\{h_{ab}\}\,d^3x$ is the volume element of Σ. From (17.68), we see that n has zero divergence iff K has zero trace, i. e. the slices are of zero mean curvature. The condition on α for this to be preserved under evolution follows from

$$0 = L_n(h^{ab}K_{ab}) = -K^{ab}L_nh_{ab} + h^{ab}L_nK_{ab}\,. \tag{17.139}$$

Here we use (17.69) to eliminate L_nh_{ab} in the first term and (17.118) to eliminate L_nK_{ab} in the second term, also making use of (17.122). This leads to the following equivalent of (17.139)

$$\Delta_h\alpha + \varepsilon\left(\mathbf{Ric}(n, n) + K^{ab}K_{ab}\right)\alpha = 0\,. \tag{17.140}$$

This is a linear elliptic equation for α. The case of interest to us in GR is $\varepsilon = -1$. In the closed case we immediately deduce by standard arguments that $\alpha = 0$ is

the only solution, provided the strong energy condition holds (which implies that $\mathbf{Ric}(n,n) \geq 0$). In the open case, where we might impose $\alpha \to 1$ as asymptotic condition, we deduce existence and uniqueness again under the assumption of the strong energy condition. Hence, we may indeed impose the condition $h^{ab}K_{ab} = 0$, or $\mathrm{Tr}(\mathbf{Wein}) = 0$, for nonclosed Σ. It is called the *maximal slicing condition* or *York gauge* [17.34].

Whereas this gauge condition has indeed the desired singularity-avoiding character it is also not easy to implement due to the fact that at each new stage of the evolution one has to solve the elliptic equation (17.140). For numerical studies it is easier to implement evolution equations for α. Such an equation is, e.g., obtained by asking the time function (17.25) to be harmonic, in the sense that

$$\begin{aligned} 0 = \Box_g t &:= g^{\mu\nu}\nabla_\mu \nabla_\nu t \\ &= |\det\{g_{\mu\nu}\}|^{-\frac{1}{2}}\partial_\mu \left(|\det\{g_{\mu\nu}\}|^{\frac{1}{2}} g^{\mu\nu}\partial_\nu \right) t . \end{aligned} \tag{17.141}$$

This is clearly just equivalent to

$$\partial_\mu \left(|\det\{g_{\mu\nu}\}|^{\frac{1}{2}} g^{\mu 0} \right) = 0 , \tag{17.142}$$

which can be rewritten using (17.54) and (17.55) to give

$$\begin{aligned} \dot\alpha := \frac{\partial \alpha}{c\partial t} &= L_\beta \alpha - \varepsilon K^a_a \alpha^2 \\ &= L_\beta \alpha + \mathrm{Tr}(\mathbf{Wein})\,\alpha^2 . \end{aligned} \tag{17.143}$$

This is called the *harmonic slicing condition*. Note that we can still choose $\beta = 0$ and try to determine α as a function of the trace of **Wein**. There also exist generalizations to this condition where α^2 on the right-hand side is replaced with other functions $f(\alpha)$.

17.6.2 A Note on the DeWitt Metric

At each point p on Σ the DeWitt metric (17.106) can be regarded as a symmetric bilinear form on the space of positive-definite inner products h of $T_p\Sigma$. The latter is an open convex cone in $T_p^*\Sigma \otimes T_p^*\Sigma$. We wish to explore its properties a little further.

A frame in $T_p\Sigma$ induces a frame in $T_p^*\Sigma \otimes T_p^*\Sigma$ (tensor product of the dual frame). If h_{ab} are the components of h, then we have the following representation of the generalized DeWitt metric

$$G_{(\lambda)} = G^{abcd}_{(\lambda)}\, \mathrm{d}h_{ab} \otimes \mathrm{d}h_{cd} , \tag{17.144a}$$

where

$$G^{abcd}_{(\lambda)} = \frac{1}{2}(h^{ac}h^{bd} + h^{ad}h^{bc} - 2\lambda h^{ab}h^{cd}) . \tag{17.144b}$$

Here we introduced a factor λ in order to parameterize the impact of the negative trace term. We also consider Σ to be of general dimension n.

The inverse metric to (17.144) is given by

$$G^{-1}_{(\lambda)} = G^{(\lambda)}_{abcd} \frac{\partial}{\partial h_{ab}} \otimes \frac{\partial}{\partial h_{cd}} , \tag{17.145a}$$

where

$$G^{(\lambda)}_{abcd} = \tfrac{1}{2}\left(h_{ac}h_{bd} + h_{ad}h_{bc} - 2\mu h_{ab}h_{cd}\right) . \tag{17.145b}$$

The relation between λ and μ is

$$\lambda + \mu = n\lambda\mu , \tag{17.146}$$

so that

$$G^{abnm}_{(\lambda)} G^{(\lambda)}_{nmcd} = \tfrac{1}{2}\left(\delta^a_c\delta^b_d + \delta^a_d\delta^b_c\right) . \tag{17.147}$$

If we change coordinates according to

$$\begin{aligned} \tau &:= \ln\left([\det\{h_{ab}\}]^{\frac{1}{n}}\right) , \\ r_{ab} &:= \frac{h_{ab}}{[\det\{h_{ab}\}]^{\frac{1}{n}}} , \end{aligned} \tag{17.148}$$

where τ parameterizes conformal changes and r_{ab} the conformally invariant ones, the metric (17.144) reads

$$G_{(\lambda)} = n(1-\lambda n)\,\mathrm{d}\tau \otimes \mathrm{d}\tau + r^{ac}r^{bd}\,\mathrm{d}r_{ab} \otimes \mathrm{d}r_{cd} , \tag{17.149}$$

where $r^{am}r_{mb} = \delta^a_b$. Since h is positive definite, so is r. Hence, the second part is positive definite on the $(\frac{1}{2}n(n+1)-1)$-dimensional vector space of trace-free symmetric tensors. Hence, the DeWitt metric is positive definite for $\lambda < 1/n$, Lorentzian for $\lambda > 1/n$, and simply degenerate (one-dimensional null space) for the critical value $\lambda = 1/n$. In the GR case we have $\lambda = 1$ and $n = 3$, so that the DeWitt metric is Lorentzian of signature $(-,+,+,+,+,+)$. Note that this Lorentzian signature is independent of ε, i. e. it has nothing to do with the Lorentzian signature of the spacetime metric.

In the Hamiltonian formulation it is not G but rather a conformally related metric that is important, the conformal factor being $\sqrt{\det\{h_{ab}\}}$. If we set

$$\hat{G}_{(\lambda)} := [\det\{h_{ab}\}]^{1/2}\, G_{(\lambda)}\,, \tag{17.150}$$

and correspondingly

$$\hat{G}^{-1}_{(\lambda)} := [\det\{h_{ab}\}]^{-1/2}\, G^{-1}_{(\lambda)}\,, \tag{17.151}$$

we can again write $\hat{G}_{(\lambda)}$ in terms of (τ, r_{ab}). In fact, the conformal rescaling clearly just corresponds to multiplying (17.149) with $\sqrt{\det\{h_{ab}\}} = e^{n\tau/2}$. Setting

$$T := 4\left[\frac{1-n\lambda}{n}\right]^{1/2} e^{n\tau/4}\,, \tag{17.152}$$

we get, excluding the degenerate case $\lambda \neq 1/n$,

$$\hat{G}_{(\lambda)} = \mathrm{sign}(1-n\lambda)\,\mathrm{d}T\otimes\mathrm{d}T + T^2 C r^{ac} r^{bd}\,\mathrm{d}r_{ab}\otimes\mathrm{d}r_{cd}\,, \tag{17.153}$$

where $C = n/(16|1-n\lambda|)$ $(=3/32$ in GR). This is a simple warped-product metric of $\mathbb{R}_+$ with the left-invariant metric on the homogeneous space $GL(3,\mathbb{R})/SO(3)\times\mathbb{R}_+$ of symmetric positive-definite forms modulo overall scale, the warping function being just T^2 if T is the coordinate on $\mathbb{R}_+$. Now, generally, quadratic warped-product metrics of the form $\pm\mathrm{d}T\otimes\mathrm{d}T + T^2 g$, where g is independent of T, are nonsingular for $T \searrow 0$ iff g is a metric of constant curvature ± 1 (as for a unit sphere in $\mathbb{R}^n$, with T being the radius coordinate, or the unit spacelike hyperboloid in n-dimensional Minkowski space, respectively). This is not the case for (17.153), which therefore has a curvature singularity for small T, i. e. small $\det\{h_{ab}\}$. Note that this is a singularity in the space of metrics (here at a fixed space point), which has nothing to do with spacetime singularities. In the early days of canonical quantum gravity this has led to speculations concerning *natural* boundary conditions for the wave function, whose domain is the space of metrics [17.35, 36]. The intention was to pose conditions such that the wave function should stay away from such singular regions in the space of metrics; see also [17.37] for a more recent discussion.

We stress once more that the signature of the DeWitt metric is not related to the signature of spacetime (it is independent of ε). For example, for the GR values $\lambda = 1$ and $n = 3$, it is Lorentzian even if spacetime were given a Riemannian metric. Moreover, by integrating over Σ, the pointwise metric (17.153) defines a bilinear form on the infinite-dimensional space of Riemannian structures on Σ, the geometry of which may be investigated to some limited extent [17.38, 39].

17.7 Constrained Hamiltonian Systems

In this section we wish to display some characteristic features of Hamiltonian dynamical systems with constraints. We restrict attention to finite-dimensional systems in order to not overload the discussion with analytical subtleties.

Let Q be the n-dimensional configuration manifold of a dynamical system that we locally coordinatize by $(q^1,\dots,q^n)$. By TQ we denote its tangent bundle, which we coordinatize by $(q^1,\dots,q^n,\, v^1,\dots,v^n)$, so that a tangent vector $X\in TQ$ is given by $X = v^a\partial/\partial q^a$. The dynamics of the system is described by a *Lagrangian*

$$L: TQ \to \mathbb{R}\,, \tag{17.154}$$

which selects the dynamically possible trajectories in TQ as follows: let $\mathbb{R}\ni t\mapsto x(t)\in Q$ be a (at least twice continuously differentiable) curve; then it is dynamically possible iff the following *Euler–Lagrange equations* hold (we set $\mathrm{d}x/\mathrm{d}t =: \dot{x}$)

$$\left.\frac{\partial L}{\partial q^a}\right|_{\substack{q=x(t)\\ v=\dot{x}(t)}} - \frac{\mathrm{d}}{\mathrm{d}t}\left[\left.\frac{\partial L}{\partial v^a}\right|_{\substack{q=x(t)\\ v=\dot{x}(t)}}\right] = 0\,. \tag{17.155}$$

Performing the t-differentiation on the second term, this is equivalent to

$$H_{ab}\left(x(t),\dot{x}(t)\right)\ddot{x}^b = V_a\left(x(t),\dot{x}(t)\right)\,, \tag{17.156}$$

where

$$H_{ab}(q,v) := \frac{\partial^2 L(q,v)}{\partial v^a\partial v^b} \tag{17.157}$$

and

$$V_a(q,v) := \frac{\partial L(q,v)}{\partial q^a} - \frac{\partial^2 L(q,v)}{\partial v^a\partial q^b}\, v^b\,. \tag{17.158}$$

Here we regard H and V as functions on TQ with values in the symmetric $n \times n$ matrices and $\mathbb{R}^n$, respectively. In order to be able to solve (17.156) for the second derivative $\ddot{x}$, the matrix H has to be invertible, that is, it must have rank n. That is the case usually encountered in mechanics. On the other hand, *constrained systems* are those where the rank of H is not maximal. This is the case we are interested in.

We assume H to be of constant rank $r < n$. Then, for each point on TQ, there exist $s = n - r$ linearly independent kernel elements $K_{(\alpha)}(q, v)$, $\alpha = 1, \ldots, s$, such that $K^a_{(\alpha)}(q, v) H_{ab}(q, v) = 0$. Hence, any solution $x(t)$ to (17.156) must be such that the curve $t \mapsto (x(t), \dot{x}(t))$ in TQ stays on the subset

$$C := \{(q, v) \in TQ \mid \psi_\alpha(q, v) = 0\,,\ \alpha = 1, \ldots, s\}\,, \tag{17.159a}$$

where

$$\psi_\alpha(q, v) = K^a_{(\alpha)}(q, v) V_a(q, v)\,. \tag{17.159b}$$

We assume $C \subset TQ$ to be a smooth closed submanifold of codimension s, i. e. of dimension $2n - s = n + r$.

Now we consider the cotangent bundle T^*Q over Q, which we coordinatize by $(q^1, \ldots, q^n, p_1, \ldots, p_n)$, so that a covector $\theta \in T^*Q$ is given by $\theta = p_a \mathrm{d}q^a$. The Lagrangian defines a map $\mathrm{FL} : TQ \to T^*Q$ through

$$\mathrm{FL}(q, v) = \left(q, p := \frac{\partial L(q, v)}{\partial v}\right)\,. \tag{17.160}$$

From what has been said above, it follows that the Jacobian of that map has constant rank $n + r$. Given sufficient regularity, we may further assume that

$$C^* := \mathrm{FL}\,(C) \subset T^*Q \tag{17.161}$$

is a smoothly embedded closed submanifold in phase space T^*Q of codimension s. Hence, there are s functions ϕ_α, $\alpha = 1, \ldots, s$ such that

$$C^* := \left\{(q, p) \in T^*Q \mid \phi_\alpha(q, p) = 0\,,\ \alpha = 1, \ldots, s\right\}\,. \tag{17.162}$$

This is called the *constraint surface* in phase space. It is given as the intersection of the zero-level sets of s independent functions. Independence means that at each $p \in C^*$ the s 1-forms $\mathrm{d}\phi_1|_p, \ldots, \mathrm{d}\phi_s|_p$ are linearly independent elements of $T^*_pT^*Q$.

The dynamical trajectories of our system will stay entirely on C^*. The trajectories themselves are integral lines of a Hamiltonian flow. But what is the Hamiltonian function that generates this flow? To explain this, we first recall the definition of the *energy function* for the Lagrangian L. It is a function $E : TQ \to \mathbb{R}$ defined through

$$E(q, v) := \frac{\partial L(q, v)}{\partial v^a} v^a - L(q, v)\,. \tag{17.163}$$

At first sight this function cannot be defined on phase space, for we cannot invert FL to express v as a function of q and p which we could insert into $E(q, v)$ in order to get $E(q, v(q, p))$. However, one may prove the following: there exists a function

$$H_{C^*} : C^* \to \mathbb{R}\,, \tag{17.164a}$$

so that

$$E = H_{C^*} \circ \mathrm{FL}\,. \tag{17.164b}$$

A local version of this is seen directly from taking the differential of (17.163), which yields $\mathrm{d}E = v^a\,\mathrm{d}(\partial L/\partial v^a) - (\partial L/\partial q^a)\,\mathrm{d}q^a$, expressing the fact that $\mathrm{d}E(X) = 0$ if $\mathrm{dFL}(X) = 0$.

So far the function H_{C^*} is only defined on C^*. By our regularity assumptions there exists a smooth extension of it to T^*Q, that is, a function $H_0 : T^*Q \to \mathbb{R}$ such that $H_0|_{C^*} = H_{C^*}$. This is clearly not unique. But, we can state the following: let H_0 and H both be smooth (at least continuously differentiable) extensions of H_{C^*} to T^*Q; then there exist s smooth functions $\lambda^\alpha : T^*Q \to \mathbb{R}$ such that

$$H = H_0 + \lambda^\alpha \phi_\alpha\,. \tag{17.165}$$

Locally a proof is simple: let $f : T^*Q \to \mathbb{R}$ be continuously differentiable and such that $f|_{C^*} \equiv 0$. Consider a point $p \in C^*$ and coordinates $(x^1, \ldots, x^{2n-s}, y^1, \ldots y^s)$ in a neighborhood $U \subset T^*Q$ of p, where the x's are coordinates on the constraint surface and the y's are just the functions ϕ. In U the constraint surface is clearly just given by $y^1 = \ldots = y^s = 0$. Then

$$\begin{aligned} f|_U(x, y) &= \int_0^1 \mathrm{d}t \frac{\mathrm{d}}{\mathrm{d}t} f(x, ty) \\ &= \int_0^1 \mathrm{d}t \frac{\partial f}{\partial y^\alpha}(x, ty)\, y^\alpha = \lambda_\alpha(x, y) y^\alpha\,, \end{aligned} \tag{17.166a}$$

where

$$\lambda_\alpha(x,y) := \int_0^1 \mathrm{d}t \frac{\partial f}{\partial y^\alpha}(x,ty)\,. \tag{17.166b}$$

For a global discussion, see [17.5].

As *Hamiltonian* for our constraint system we address any smooth (at least continuously differentiable) extension H of H_{C^*}. So, if H_0 is a somehow given one, any other can be written as

$$H = H_0 + \lambda^\alpha \phi_\alpha\,, \tag{17.167}$$

for some (at least continuously differentiable) real-valued functions λ^α on T^*Q.

Here we have been implicitly assuming that the Hamiltonian dynamics does not leave the constraint surface (17.161). If this were not the case, we would have to restrict further to proper submanifolds of C^* such that the Hamiltonian vector fields evaluated on them lie tangentially. (If no such submanifold can be found, the theory is simply empty.) This is sometimes expressed by saying that the *primary constraints* (those encountered first in the Lagrangian/Hamiltonian analysis) are completed by *secondary*, *tertiary*, etc. constraints for consistency.

Here we assume that our system is already dynamically consistent. This entails that the Hamiltonian vector fields X_{ϕ_α} for the ϕ_α are tangential to the constraint surface. This is equivalent to $X_{\phi_\alpha}(\phi_\beta)|_{C^*} = 0$, or expressed in Poisson brackets

$$\{\phi_\alpha, \phi_\beta\}|_{C^*} = 0\,, \tag{17.168}$$

for all $\alpha, \beta \in \{1, \dots, s\}$. Following *Dirac* [17.3], constraints which satisfy this condition are said to be of *first class*. By the result shown (locally) above in (17.166), this is equivalent to the existence of $\frac{1}{2}s^2(s-1)$ (at least continuously differentiable) real-valued functions $C^\gamma_{\alpha\beta} = -C^\gamma_{\beta\alpha}$ on T^*Q, such that

$$\{\phi_\alpha, \phi_\beta\} = C^\gamma_{\alpha\beta}\phi_\gamma\,. \tag{17.169}$$

Note that, as far as the intrinsic geometric properties of the constraint surface are concerned, (17.168) and (17.169) are equivalent.

The indeterminacy of the Hamiltonian due to the freedom to choose any set of λ^α seems to imply an s-dimension worth of indeterminacy in the dynamically allowed motions. But, the difference in these motions is that generated by the constraint functions on the constraint surface. In order to actually tell apart two such motions requires observables (phase-space functions) whose Poisson brackets with the constraints do *not* vanish on the constraint surface. The general attitude is to assume that this is not possible, i. e. to assume that *physical observables* correspond exclusively to phase-space functions whose Poisson brackets with all constraints vanish on the constraint surface. This is expressed by saying that all motions generated by the constraints are *gauge transformations*. This entails that they are undetectable in principle and merely correspond to a mathematical redundancy in the description rather than to any physical degrees of freedom. It is therefore more correct to speak of gauge *redundancies* rather than of gauge *symmetries*, as is sometimes done, for the word *symmetry* is usually used for a physically meaningful operation that does change the object to which it is applied in at least some aspects (otherwise the operation is the identity). Only some *relevant* aspects, in the context of which one speaks of symmetry, are not changed.

17.7.1 Geometric Theory

Being first class has an interpretation in terms of symplectic geometry. To see this, we first recall a few facts and notation from elementary symplectic geometry. Here some sign conventions enter and the reader is advised to compare carefully with other texts.

A *symplectic structure* on a manifold is a nondegenerate closed 2-form. On any cotangent bundle there is a natural such structure which derives from a *symplectic potential*. The latter is a 1-form field θ on T^*Q whose general geometric definition is as follows: let $\pi : T^*Q \to Q$ be the natural projection from the cotangent bundle of Q (phase space) to Q itself. Then, for each $p \in T^*Q$, we define

$$\theta_p := p \circ \pi_*|_p\,. \tag{17.170}$$

So, in order to apply θ_p to a vector $X \in T_pT^*Q$, we do the following: take the differential π_* of the projection map π, evaluate it at point p, and apply it to $X \in T_pT^*Q$ in order to push it forward to the tangent space $T_{\pi(p)}Q$ at point $\pi(p) \in Q$. Then apply p to it, which makes sense since p is, by definition, an element of the cotangent space at $\pi(p) \in Q$.

In the coordinates already introduced, this form is simply given by

$$\theta := p_a \mathrm{d}q^a\,. \tag{17.171}$$

The symplectic structure is a 2-form field on T^*Q, now simply defined as

$$\omega := -\mathrm{d}\theta = \mathrm{d}q^a \wedge \mathrm{d}p_a . \tag{17.172}$$

The nondegeneracy of ω allows us to uniquely associate a vector field X_f with any real-valued function f on T^*Q through

$$i_{X_f}\omega = \mathrm{d}f . \tag{17.173}$$

It is called the *Hamiltonian vector field* of f. An immediate consequence of (17.173) and $\mathrm{d}\omega = 0$ (which follows from (17.171)) is that ω has vanishing Lie derivative with respect to any Hamiltonian vector field

$$L_{X_f}\omega = (i_{X_f} \circ d + d \circ i_{X_f})\omega = 0 . \tag{17.174}$$

In coordinates, X_f looks like this

$$X_f = \frac{\partial f}{\partial p_a}\frac{\partial}{\partial q^a} - \frac{\partial f}{\partial q^a}\frac{\partial}{\partial p_a} . \tag{17.175}$$

The *Poisson bracket* between two functions f and g is defined as

$$\begin{aligned}\{f,g\} &:= \omega(X_f, X_g) = X_g(f) = -X_f(g)\\ &= \frac{\partial f}{\partial q^a}\frac{\partial g}{\partial p_a} - \frac{\partial f}{\partial p_a}\frac{\partial g}{\partial q^a} .\end{aligned} \tag{17.176}$$

It provides $C^\infty(T^*Q)$ with a structure of a Lie algebra, which, if taken together with the commutative and associative pointwise multiplication, endows $C^\infty(T^*Q)$ with the structure of a *Poisson algebra*. The map $f \mapsto X_f$ is a Lie anti-homomorphism from the Lie algebra of functions to the Lie algebra of vector fields on T^*Q (the Lie multiplication of the latter is just the commutator). This is expressed by

$$X_{\{f,g\}} = -[X_f, X_g] , \tag{17.177}$$

which is easy to prove from the definitions given.

After this brief digression, we now return to the geometric interpretation of first-class constraints. For any $p \in C^*$, we define

$$\begin{aligned}&T_p^\perp(T^*Q)\\ &:= \left\{X \in T_p(T^*Q) \mid \omega(X,Y) = 0 , \forall Y \in T_pC^*\right\} .\end{aligned} \tag{17.178}$$

The nondegeneracy of ω implies that the dimension of $T_p^\perp(T^*Q)$ equals s, the codimension of C^* in T^*Q. But, note that as ω is skew, $T_p^\perp(T^*Q)$ might well have a nontrivial intersection with T_pC^*. This gives rise to the following characterizations for the submanifold $C^* \subset T^*Q$ (understood to hold at each point $p \in C^*$): C^* is called:

- *Isotropic* iff $T_pC^* \subset T_p^\perp(T^*Q)$;
- *Co-isotropic* iff $T_pC^* \supset T_p^\perp(T^*Q)$;
- *Lagrangian* iff $T_pC^* = T_p^\perp(T^*Q)$.

Since $\{\phi_\alpha, \phi_\beta\} = \mathrm{d}\phi_\alpha(X_{\phi_\beta})$, we see that condition (17.168) is equivalent to the statement that the Hamiltonian vector fields for the constraint functions ϕ_α are tangent to the constraint hypersurface

$$X_{\phi_\alpha}|_{C^*} \in \Gamma TC^* . \tag{17.179}$$

Our assumption that the s differentials $\mathrm{d}\phi_\alpha$ be linearly independent at each $p \in C^*$ now implies that the s vectors $X_{\phi_\alpha}(p)$ span an s-dimensional subspace of T_pC^*. But, they are also elements of $T_p^\perp(T^*Q)$, since $\omega(X_{\phi_\alpha}, Y) = \mathrm{d}\phi_\alpha(Y) = 0$ for all Y tangent to C^*. As the dimension of $T_p^\perp(T^*Q)$ is s, this shows that

$$T_p^\perp(T^*Q) = \mathrm{span}\{X_{\phi_1}, \dots, X_{\phi_s}\} \subset T_pC^* , \tag{17.180}$$

that is, co-isotropy of C^*. First-class constraints are precisely those which give rise to co-isotropic constraint surfaces.

The significance of this lies in the following result, which we state in an entirely intrinsic geometric fashion. Let C^* be co-isotropic of codimension s and let $e : C^* \to T^*Q$ be its embedding. We write

$$\hat{\omega} := e^*\omega \tag{17.181}$$

for the pull back of ω to the constraint surface (i. e. essentially the restriction of ω to the tangent bundle of the constraint surface). $\hat{\omega}$ is now s-fold degenerate, its kernel at $p \in C^*$ being just $T_p^\perp(T^*Q) \subset T_pC^*$. We have the smooth assignment of subspaces

$$C^* \ni p \mapsto \mathrm{kernel}_p(\hat{\omega}) = T_p^\perp(T^*Q) , \tag{17.182}$$

which forms a subbundle of TC^* called the *kernel distribution* of $\hat{\omega}$. Now, the crucial result is that this subbundle is *integrable*, i. e. tangent to locally embedded submanifolds $\gamma^* \subset C^*$ of codimension s in C^*, or codimension $2s$ in T^*Q. Indeed, in order to show this

we only need to show that whenever two vector fields X and Y on C^* take values in the kernel distribution their commutator $[X, Y]$ also takes values in the kernel distribution. That this suffices for local integrability is known as *Frobenius' theorem* in differential geometry. Writing

$$i_{[X,Y]}\hat{\omega} = L_X\left(i_Y\hat{\omega}\right) - i_Y\left(L_X\hat{\omega}\right), \tag{17.183}$$

we infer that the first term vanishes because X is in $\hat{\omega}$'s kernel and $L_X\hat{\omega}$ vanishes because $L_X = \mathrm{d}\circ i_X + i_X\circ\mathrm{d}$ on forms, where $i_X\hat{\omega} = 0$ again due to X being in the kernel and $\mathrm{d}\hat{\omega} = \mathrm{d}e^*\omega = e^*\,\mathrm{d}\omega = 0$ due to ω being closed.

The program of *symplectic reduction* is now to form the $(2n-2s)$-dimensional quotient space $C^*/\sim$, where $\sim$ is the equivalence relation whose equivalence classes are the maximal integral submanifolds of the kernel distribution of $\hat{\omega}$. We stress that this geometric formulation of the reduction program does not refer to any set of functions ϕ_α that one might use in order to characterize C^*. If one uses such functions, it is understood that they obey the above-mentioned regularity conditions of being at least continuously differentiable in a neighborhood of C^* and giving rise to a set of s linearly independent differentials $\mathrm{d}\phi_\alpha$ at any point of C^*. Hence, redefinitions of constraint functions like $\phi\mapsto\sqrt{|\phi|}$ or $\phi\mapsto\phi^2$, albeit leading to the same surface C^*, are a priori not allowed.

17.7.2 First-Class Constraints from Zero-Momentum Maps

First-class constraints often arise from group actions. This is also true in GR, at least partially. So, let us explain this in more detail. Let a Lie group G act on the left on T^*Q. This means that there is a map $G\times T^*Q\to T^*Q$, denoted simply by $(g,p)\mapsto g\cdot p$, so that $g_1\cdot(g_2\cdot p) = (g_1g_2)\cdot p$ and $e\cdot p = p$ if $e\in G$ is the neutral element. There is then an anti-homomorphism from Lie(G), the Lie algebra of G, to the Lie algebra of vector fields on T^*Q. It is defined as follows: the vector field V_ξ corresponding to $\xi\in\mathrm{Lie}(G)$, evaluated at point $p\in T^*Q$, is

$$V_\xi(p) := \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0}\exp(t\xi)\cdot p\,. \tag{17.184}$$

It is then not hard to prove that

$$\left[V_\xi, V_\eta\right] = -V_{[\xi,\eta]}\,. \tag{17.185}$$

Let us further suppose that the group action on T^*Q is of a special type, namely it arises from a group action on Q by a canonical lift. (Every diffeomorphism f of Q can be lifted to a diffeomorphism $\hat{f}$ of T^*Q given by the pull back of the inverse f^{-1}.) Then it is easy to see from the geometric definition (17.170) that the symplectic potential θ is invariant under this group action and consequently the group acts by symplectomorphisms (ω-preserving diffeomorphisms). The infinitesimal version of this statement is that, for all $\xi\in\mathrm{Lie}(G)$

$$L_{V_\xi}\theta = 0\,. \tag{17.186}$$

Since $L_{V_\xi} = i_{V_\xi}\circ\mathrm{d} + \mathrm{d}\circ i_{V_\xi}$, this is equivalent to

$$i_{V_\xi}\omega = \mathrm{d}\left(\theta(V_\xi)\right)\,, \tag{17.187}$$

which says that V_ξ is the Hamiltonian vector field of the function $\theta(V_\xi)$. We call the map

$$\mathrm{Lie}(G)\ni\xi\mapsto P_\xi := \theta(V_\xi)\in C^\infty(T^*Q) \tag{17.188}$$

the *momentum map* for the action of G. It is a linear map from Lie(G) to $C^\infty(T^*Q)$ and satisfies

$$\begin{aligned}\{P_\xi, P_\eta\} &= V_\eta\left(\theta(V_\xi)\right)\\ &= \left(L_{V_\eta}\theta\right)(V_\xi) + \theta\left(L_{V_\eta}V_\xi\right)\\ &= \theta\left(V_{[\xi,\eta]}\right)\\ &= P_{[\xi,\eta]}\,,\end{aligned} \tag{17.189}$$

where we used (17.186) and (17.185) for the third equality. Hence, we see that the map (17.188) is a Lie homomorphism from Lie(G) into the Lie algebra of smooth, real-valued functions on T^*Q (whose Lie product is the Poisson bracket).

Now, first-class constraints are often given by the condition of *zero-momentum mappings*, i. e., by $P_\xi = 0$ for all $\xi\in\mathrm{Lie}(G)$. By linearity in ξ, this is equivalent to the set of $s := \dim(G)$ conditions

$$\phi_\alpha := P_{e_\alpha} = 0\,, \tag{17.190}$$

where $e_\alpha = \{e_1,\dots,e_s\}$ is a basis of Lie(G). Let the structure constants for this basis be $C^\gamma_{\alpha\beta}$, i. e. $[e_\alpha, e_\beta] = C^\gamma_{\alpha\beta}e_\gamma$; then (17.189) becomes

$$\{\phi_\alpha, \phi_\beta\} = C^\gamma_{\alpha\beta}\phi_\gamma\,. \tag{17.191}$$

Constraints in gauge theories will typically arise as zero-momentum maps in the fashion described here, the only necessary generalization being the extension to infinite-dimensional groups and Lie algebras. In fact, for gauge theories our G will correspond to the infinite-dimensional *group of gauge transformations*, which is not to be confused with the finite-dimensional gauge group. The former consists of functions, or sections in bundles, with values in the latter. On the other hand, the constraints in GR will only partially be of this type. More precisely, those constraints arising from three-dimensional diffeomorphisms (called the vector or diffeomorphism constraints) will be of this type; those from nontangential hypersurface deformations (scalar or Hamiltonian constraints) will not fit into this picture. For the former, G will correspond to $\mathrm{Diff}(\Sigma)$, or some appropriate subgroup thereof, and $\mathrm{Lie}\,(\mathrm{Diff}(\Sigma))$ to the infinite-dimensional Lie algebra of vector fields on Σ (possibly with special support and/or fall-off conditions). The different nature of the latter constraint will be signaled by structure *functions* $C^{\gamma}_{\alpha\beta}(q,p)$ appearing on the right-hand side rather than constants. This has recently given rise to attempts to generalize the group-theoretic setting described above to that of *groupoids* and *Lie algebroids*, in which the more general structure of GR can be accommodated [17.40].

17.8 Hamiltonian GR

The Hamiltonian formulation of GR proceeds along the lines outlined in the previous section. For this we write down the action in a $(3+1)$-split form, read off the Lagrangian density, defining the conjugate momenta as derivatives of the latter with respect to the velocities, and finally expressing the energy function (17.163) in terms of momenta. The constraint functions will not be determined on the Lagrangian level, but rather directly on the Hamiltonian level as primary and secondary constraints (there will be no tertiary ones), the primary ones being just the vanishing of the momenta for lapse and shift.

The Lagrangian density for GR is essentially just the scalar curvature of spacetime. However, upon variation of this quantity, which contains second derivatives in the metric, we will pick up boundary terms from partial integrations which need not vanish by just keeping the metric on the boundary fixed. Hence, we will need to subtract these boundary terms, which will otherwise obstruct functional differentiability. Note that this is not just a matter of aesthetics: solutions to differential equations (like Einstein's equations) will not be stationary points of the action if the latter is not differentiable at these points. Typically, Euler–Lagrange equations will allow for solutions outside the domain of differentiability of the action they are derived from. Including some such solutions will generally need the adaptation of the action by boundary terms. This clearly matters if one is interested in the values of the action, energies, etc. for these solutions and, also, of course, in the path-integral formulations of the corresponding quantum theories.

The Einstein–Hilbert action of GR is

$$S_{\mathrm{GR}}[\Omega, g] = -\frac{\varepsilon}{2\kappa}\int_{\Omega} \mathbf{Scal}\,\mathrm{d}\mu_g + \text{boundary terms}\,, \tag{17.192}$$

where, in local coordinates $x^{\mu} = (x^0 = ct, x^1, x^2, x^3)$

$$\mathrm{d}\mu_g = \sqrt{\varepsilon \det\{g_{\mu\nu}\}}c\,\mathrm{d}t \wedge \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3\,. \tag{17.193}$$

The sign convention behind the prefactor $-\varepsilon$ in (17.192) is such that in the Lorentzian as well as the Riemannian case the Lagrangian density contains the bilinear DeWitt inner product of the extrinsic curvatures (compare (17.128)) with a positive sign, i. e. transverse traceless modes have positive kinetic energy.

The boundary term can be read off from (17.128) and (17.126). If the integration domain $\Omega \subset M$ is such that the spacelike boundaries are contained in two hypersurfaces Σ_s, i. e. two $t=$ const. surfaces, say $\partial\Omega_i := \partial\Omega \cap \Sigma_{\text{initial}}$ and $\partial\Omega_f := \partial\Omega \cap \Sigma_{\text{final}}$, we would have to add the two boundary terms (dependence on ε drops out)

$$\kappa^{-1}\int_{\partial\Omega_f} \mathrm{Tr}_h(K)\,\mathrm{d}\mu_h - \kappa^{-1}\int_{\partial\Omega_i} \mathrm{Tr}_h(K)\,\mathrm{d}\mu_h\,. \tag{17.194}$$

Here we used that the second term in (17.126) does not contribute due to a being orthogonal to n. $\mathrm{d}\mu_h$ is the

standard measure from the induced metric h on the hypersurfaces. If the cylindrical timelike boundary $\partial\Omega_{\mathrm{cyl}}$ is chosen such that its spacelike normal m is orthogonal to n, only the second term in (17.126) contributes and we get one more boundary term (again ε drops out)

$$\kappa^{-1}\int_{\partial\Omega_{\mathrm{cyl}}}\hat{K}(n,n)\,\mathrm{d}\mu_h\,. \tag{17.195}$$

Here $\hat{K}$ is the extrinsic curvature of $\partial\Omega_{\mathrm{cyl}}$ in M, which we picked up because $g(a,m)=g(\nabla_n n,m)=-g(n,\nabla_n m)=\hat{K}(n,n)$.

Once the boundary terms are taken care of, we can just read off the Lagrangian density from (17.128) also using (17.55)

$$\mathcal{L}_{\mathrm{GR}}=(2\kappa)^{-1}\left[G(K,K)-\varepsilon R\right]\alpha\sqrt{h}\,, \tag{17.196}$$

where we now used the standard abbreviations

$$\begin{aligned}G(K,K)&:=G^{abcd}K_{ab}K_{cd}\,,\\ R&:=\mathbf{Scal}^D\,,\\ \sqrt{h}&:=\sqrt{\det\{h_{ab}\}}\,.\end{aligned} \tag{17.197}$$

Moreover, K_{ab} has here to be understood as expressed in terms of the time and Lie derivatives of h_{ab}

$$K=-\frac{\varepsilon}{2}\alpha^{-1}\left(\dot{h}-L_\beta h\right)\,. \tag{17.198}$$

We keep in mind that an overdot denotes differentiation with respect to ct (not t). In passing, we also note that $\mathcal{L}_{\mathrm{GR}}$ has the right physical dimension of an energy density (α is dimensionless).

The Hamiltonian density is now obtained by the usual Legendre transform with respect to all configuration variables that are varied in the action. These comprise all components $g_{\mu\nu}$ and hence in the $(3+1)$-split parameterization all h_{ab} as well as the lapse α and the three shift components β^a. However, it is immediate that (17.196) does not contain any time derivatives of the latter; hence, their conjugate momenta vanish

$$\pi_\alpha:=\frac{1}{c}\frac{\partial\mathcal{L}_{\mathrm{GR}}}{\partial\dot{\alpha}}=0\,, \tag{17.199a}$$

$$\pi_{\beta^a}:=\frac{1}{c}\frac{\partial\mathcal{L}_{\mathrm{GR}}}{\partial\dot{\beta}^a}=0\,. \tag{17.199b}$$

This leaves us with the momenta for the metric components h_{ab}

$$\begin{aligned}\pi^{ab}:&=\frac{1}{c}\frac{\partial\mathcal{L}_{\mathrm{GR}}}{\partial\dot{h}_{ab}}\\ &=\frac{(-\varepsilon)\sqrt{h}}{2\kappa c}G^{abcd}K_{cd}\\ &=\frac{(-\varepsilon)}{2\kappa c}\hat{G}^{abcd}K_{cd}\,.\end{aligned} \tag{17.200}$$

Here again K stands for the expression (17.198). We also made use of the conformally rescaled DeWitt metric (17.150), whose significance appears here for the first time. Again in passing we note that the physical dimension of π^{ab} is right, namely that of momentum per area (the dimension of K is an inverse length).

In order to compute the Hamiltonian density, we express $\dot{h}$ in terms of the momenta

$$\begin{aligned}\dot{h}_{ab}&=(L_\beta h)_{ab}-2\varepsilon\alpha K_{ab}\\ &=(D_a\beta_b+D_b\beta_a)-2\varepsilon\alpha K_{ab}\,,\end{aligned} \tag{17.201}$$

and obtain

$$\begin{aligned}\mathcal{H}_0[h,\pi]&=\pi^{ab}c\dot{h}_{ab}-\mathcal{L}_{\mathrm{GR}}\\ &=\alpha\Big[(2\kappa c^2)\hat{G}^{-1}_{abcd}\pi^{ab}\pi^{cd}\\ &\quad+\varepsilon(2\kappa)^{-1}\sqrt{h}R\Big]\\ &\quad+2c\pi^{ab}D_a\beta_b\,.\end{aligned} \tag{17.202}$$

The Hamiltonian H_0 is just the integral of this density over Σ. The subscript 0 is to indicate that this Hamiltonian is still to be modified by constraints according to the general scheme. Also, we have to once more care about surface terms in order to ensure functional differentiability, without which the Hamiltonian flow does not exist [17.41].

The first thing to note is that we have found the primary constraints (17.199). For them to be maintained under the evolution, we impose

$$c\dot{\pi}_\alpha=\{\pi_\alpha,H_0\}=-\frac{\delta H_0}{\delta\alpha}=0\,, \tag{17.203a}$$

$$c\dot{\pi}_{\beta^a}=\{\pi_{\beta^a},H_0\}=-\frac{\delta H_0}{\delta\beta^a}=0\,, \tag{17.203b}$$

giving rise to the secondary constraints

$$(2\kappa c^2)\hat{G}^{-1}(\pi,\pi)+\varepsilon(2\kappa)^{-1}\sqrt{h}R=0\,, \tag{17.204a}$$

$$-2D_a\pi^{ab}=0\,, \tag{17.204b}$$

respectively. It may be checked directly that these equations respectively are equivalent to (17.131) for $\mathbf{T} = 0$. If we had included a cosmological constant, this would have led to the replacement of R in (17.204a) with $(R - 2\Lambda)$. We know from our previous discussion that the requirement of these secondary constraints to be preserved in time will not lead to further constraints. This has been shown by the argument following (17.137).

The primary constraints are taken care of by simply eliminating the canonical pairs (α, π_α) and (β^a, π_{β^a}) from the list of canonical variables. As we will see shortly, the secondary constraints (17.204) are of first class, so that, according to the general theory outlined above, they should be added with arbitrary coefficients to the initial Hamiltonian H_0 to get the general Hamiltonian. This leads to

$$H[\alpha, \beta] = C_s(\alpha) + C_v(\beta) + \text{boundary terms}\,, \tag{17.205}$$

where

$$C_s(\alpha) := \int_\Sigma \mathrm{d}^3x\alpha \left[(2\kappa c^2)\hat{G}^{-1}(\pi,\pi) + \varepsilon(2\kappa)^{-1}\sqrt{h}R\right], \tag{17.206a}$$

$$C_v(\beta) := \int_\Sigma \mathrm{d}^3x\beta^a \left[-2ch_{ab}D_c\pi^{bc}\right], \tag{17.206b}$$

where α and β^a are now arbitrary coefficients corresponding to the λ's in (17.167). In particular, they may depend on the remaining canonical variables h and π. Note that up to boundary terms the Hamiltonian is just a sum of constraints, where s stands for the scalar (or Hamiltonian) and V for the vector (or diffeomorphism) constraint. The equations of motion generated by H are

$$c\dot{h}_{ab} = \{h_{ab}, H\}\,, \tag{17.207}$$

$$c\dot{\pi}^{ab} = \{\pi^{ab}, H\} \tag{17.208}$$

equivalent to (17.136) and (17.135), respectively, and need not be written down again. Before we discuss the boundary terms, we write down the Poisson brackets for the constraints

$$\{C_v(\beta), C_v(\beta')\} = C_v([\beta, \beta'])\,, \tag{17.209a}$$

$$\{C_v(\beta), C_s(\alpha)\} = C_s(\beta(\alpha))\,, \tag{17.209b}$$

$$\{C_s(\alpha), C_s(\alpha')\} = \varepsilon C_v(\alpha(\mathrm{d}\alpha')^\sharp - \alpha'(\mathrm{d}\alpha)^\sharp)\,. \tag{17.209c}$$

These may be obtained by direct computation, but are also dictated by geometry. Before discussing the geometry behind them, we note the following obvious points:

1. The vector constraints form a Lie algebra. The map $\beta \to V(\beta)$ is a Lie homomorphism from the Lie algebra of vector fields in Σ to the Lie algebra (within the Poisson algebra) of phase-space functions. In fact, this map is just the momentum map for the action of the diffeomorphism group $G = \mathrm{Diff}(\Sigma)$ on phase T^*Q, which is a lift of the action on $Q = \mathrm{Riem}(\Sigma)$, the space of Riemannian metrics on Σ. Note that here the symplectic potential can be written in a symbolic infinite-dimensional notation

$$\theta = \int_\Sigma \mathrm{d}^3x\, \pi^{ab}(x)\delta h_{ab}(x)\,, \tag{17.210}$$

and the vector field V_β generated by the action of $G = \mathrm{Diff}(\Sigma)$ on $Q = \mathrm{Riem}(\Sigma)$ as

$$V_\beta = \int_\Sigma \mathrm{d}^3x\, L_\beta h_{ab}(x)\frac{\delta}{\delta h_{ab(x)}}\,. \tag{17.211}$$

The momentum map (17.188) is then given by

$$P_\beta = \theta(V_\beta) = \int_\Sigma \mathrm{d}^3x\, \pi^{ab}L_\beta h_{ab} = c^{-1}C_v(\beta) + 2\int_{\partial\Sigma} \mathrm{d}^2x\beta_a\pi^{ab}m_b\,, \tag{17.212}$$

where m^b denote again the components of the outward-pointing normal of $\partial\Sigma$. This shows that for vector fields β for which the surface term does not contribute, the vector constraint is just the momentum map (up to a factor of c^{-1}, which comes in because the physical dimension of the values of the momentum map is that of momentum whereas the physical dimension of the constraints is that of an Hamiltonian, that is, energy). The surface term will be discussed below. What is important here is that the vector constraint coincides with the zero-momentum map for those diffeomorphisms which are asymptotically trivial, i. e. for which the surface term vanishes. Only those are to be considered as gauge transformations. Long-ranging diffeomorphisms for which the surface term is nonzero, i. e. for configurations of nonvanishing

linear and/or angular momenta (cf. Sect. 17.9), have to be considered as proper changes in physical state. If we required these motions to be pure gauge, we would eliminate all states with nonzero asymptotic charges. Compare the closing remarks of Sect. 17.7.

2. Once we have understood that the vector constraint is the momentum map for diffeomorphisms, its Poisson bracket with any other phase-space function F that defines a *geometric object* on Σ (i. e. an object with well-defined transformation properties under diffeomorphisms) is fixed. We simply have

$$\{F, V_\beta\} = L_\beta F . \tag{17.213}$$

In this sense (17.209b) says no more than that the expression (17.204a) is a scalar density of weight one. Recall that if F is a scalar density of weight one, then $L_\beta F = D_a(\beta^a F)$. If we multiply F by α and integrate over Σ, we get after partial integration and assuming the boundary term to give no contribution (which for nonclosed Σ requires certain fall-off conditions) an integral of $-F\beta(\alpha)$, which is just what (17.209b) expresses. Algebraically speaking, the fact that the Poisson bracket of a vector and a scalar constraint is proportional to a scalar rather than a vector constraint means that the vector constraints do *not* form an *ideal*. Geometrically this means that the Hamiltonian vector fields for the scalar constraint, if evaluated on the hypersurface for the vector constraint, will generally not be tangential to it, except for the points where this hypersurface intersects that of the scalar constraint. This has very important consequences for algorithms of phase-space reduction, i. e. algorithms that aim to *solve* the constraints. It means that a reduction in steps is *not* possible, whereby one first solves for the vector constraint and then seeks for solutions of the scalar constraint.

3. According to (17.209b), two scalar constraints Poisson commute into a vector constraint. Two facts are remarkable concerning the vector field that forms the argument of this vector constraint: first, it depends on the signature of spacetime (overall multiplication with ε). Second, it depends on the phase-space variable h through the $\sharp$-operation of *index raising*; explicitly

$$\begin{aligned} &\alpha(\mathrm{d}\alpha')^\sharp - \alpha'(\mathrm{d}\alpha)^\sharp \\ &= h^{ab}\left(\alpha\partial_b\alpha' - \alpha'\partial_b\alpha\right)\frac{\partial}{\partial x^a} . \end{aligned} \tag{17.214}$$

This is the fact, already mentioned at the end of Sect. 17.7, that the constraints in GR are not altogether in the form of a vanishing momentum map. This fact has led to some discussion in the past and attempts have been made to consider different algebraic combinations of the constraints which define the same constraint hypersurfaces but display structure constants rather than structure functions in their Poisson brackets; e.g., [17.42]. But, as already discussed in Sect. 17.7, it is important that these redefinitions do not spoil the regularity properties of the functions that define the constraint surface.

This ends the immediate discussion of (17.209). But there is another aspect that is related to the last point just mentioned and that deserves to be mentioned.

17.8.1 Hypersurface Deformations and Their Representations

Even though the constraints cannot be understood in a straightforward fashion as a zero-momentum map of a group action, they nevertheless do furnish a representation of an algebraic object (a groupoid) of hypersurface motions. As a result, the relations (17.209) are *universal*, in the sense that *any* spacetime diffeomorphism invariant theory, whatever its field content, will give rise to the very same relations (17.209); see [17.43, 44] for early and lucid discussions and [17.45, 46] for a comprehensive account.

The idea is to regard the space of (spacelike) embeddings $\mathrm{Emb}(\Sigma, M)$ of Σ into M as an infinite-dimensional manifold, on which the diffeomorphism group of M acts on the left by simple composition. Then there is a standard anti-homomorphism from the Lie algebra of $\mathrm{Diff}(M)$ to the Lie algebra of vector fields on $\varepsilon(\Sigma, M)$, just as in (17.185). A tangent vector at a particular $\varepsilon \in \mathrm{Emb}(\Sigma, M)$ can be visualized as a vector field ξ on $\Sigma \subset M$ with normal and tangential components; more precisely, as a section in the pull-back bundle ε^*TM over Σ. Its decomposition into normal and tangential components depends on ε. If we think of M as being locally coordinatized by functions y^μ and Σ by functions x^a, then ε can be locally represented by four functions y^μ of three variables x^a. A vector field V_ξ can then be represented in a symbolic infinite-dimensional notation

$$V_\xi = \int_\Sigma \mathrm{d}^3x\, \xi^\mu\,(y(x))\,\frac{\delta}{\delta y^\mu(x)} . \tag{17.215}$$

In full analogy to (17.185), this immediately leads to

$$[V_\xi, V_\eta] = -V_{[\xi,\eta]} \,. \tag{17.216}$$

If we now decompose ξ in an embedding-dependent fashion into its normal component αn and tangential component β, we can rewrite (17.215) as

$$V(\alpha,\beta) = \int_\Sigma \mathrm{d}^3x \, \left(\alpha(x) n^\mu[y](x) + \beta^a \partial_a y^\mu(x)\right) \frac{\delta}{\delta y^\mu(x)} \,. \tag{17.217}$$

Here $\varepsilon(\Sigma) \subset M$ has to be considered as a functional of the embedding. Again, we can compute the commutator explicitly. The only nontrivial part is the functional derivative of the n^μ with respect to the y^ν. How this is done is explained in the Appendix of [17.43]. The result is

$$[V(\alpha_1,\beta_1), V(\alpha_2,\beta_2)] = -V(\alpha,\beta) \,, \tag{17.218a}$$

where

$$\alpha = \beta_1(\alpha_2) - \beta_2(\alpha_1) \,, \tag{17.218b}$$

$$\beta = [\beta_1,\beta_2] + \varepsilon\left(\alpha_1(\mathrm{d}\alpha_2)^\sharp - \alpha_2(\mathrm{d}\alpha_1)^\sharp\right) \,. \tag{17.218c}$$

This is just (17.209) up to a relative minus sign that has the same origin as that between (17.185) and (17.189). We therefore see that (17.209) is a representation of a general algebraic structure which derives from the geometry of deformations of (spacelike) hypersurfaces in spacetime.

We can now address the inverse problem, namely to find all Hamiltonian representations of (17.218) on a given phase space. As in GR the phase space is T^*Q, where $Q = \mathrm{Riem}(\Sigma)$. That is, we may ask for the most general phase-space functions $H(\alpha,\beta)$: $T^*\mathrm{Riem}(\Sigma) \to \mathbb{R}$, parameterized by (α,β), so that

$$\{H(\alpha_1,\beta_1), H(\alpha_2,\beta_2)\} = H(\alpha,\beta) \,. \tag{17.219}$$

The meaning of this relation is once more explained in Fig. 17.3. It is also sometimes expressed as *path independence*, for it implies that the Hamiltonian flow corresponding to two different paths in $\varepsilon(\Sigma, M)$ reaching the same final hypersurface will also result in the same physical state (phase-space point).

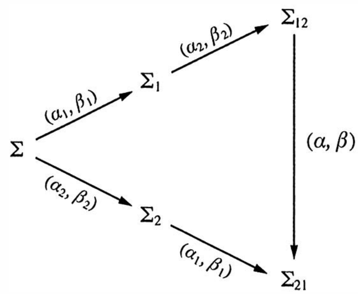


Fig. 17.3 An (infinitesimal) hypersurface deformation with parameters (α_1,β_1) that maps $\Sigma \mapsto \Sigma_1$, followed by one with parameters (α_2,β_2) that maps $\Sigma_1 \mapsto \Sigma_{12}$, differs by one with parameters (α,β) given by (17.218b) from that in which the maps with the same parameters are composed in the opposite order

To answer this question, one first has to choose a phase space. Here we stick to the same phase space as in GR, that is, T^*Q, where $Q = \mathrm{Riem}(\Sigma)$. The representation problem can be solved under certain additional hypotheses concerning the geometric interpretation of $H(\alpha = 0, \beta)$ and $H(\alpha, \beta = 0)$:

1. $H(0,\beta)$ should represent an infinitesimal spatial diffeomorphism, so that

$$\{F, H(0,\beta)\} = L_\beta F \tag{17.220a}$$

   for any phase-space function F. This fixes $H(0,\beta)$ to be the momentum map for the action of $\mathrm{Diff}(\Sigma)$ on phase space.
2. $H(\alpha, 0)$ should represent an infinitesimal $\mathrm{Diff}(M)$ action *normal to* Σ. In the absence of M, which is not yet constructed, this phrase is taken to mean that (17.69) must hold, i. e.

$$\{h, H(\alpha,0)\} = -2\varepsilon\alpha K \,, \tag{17.220b}$$

   where K is the extrinsic curvature of Σ in the ambient spacetime that is yet to be constructed.

It has been shown that under these conditions the Hamiltonian of GR, including a cosmological constant, provides the unique two-parameter family of solutions, the parameters being κ and Λ. See [17.44] for more details and [17.47] for the most complete proof (see below for a small topological gap). This result may be seen as the Hamiltonian analog to *Lovelock*'s uniqueness re-

sult [17.48] for Einstein's equations using spacetime covariance.

A particular consequence of this result is the impossibility to change the parameter λ in the DeWitt metric (17.144) to any other than the GR value $\lambda = 1$ without violating the representation condition, that is, without violating covariance under spacetime diffeomorphisms. Such theories include those of Hořava–Lifshitz type [17.49], which were suggested as candidates for ultraviolet completions of GR.

At this point, we must mention a topological subtlety which causes a small gap in the uniqueness proofs mentioned above and might have important consequences in quantum gravity. To approach this issue, we recall from the symplectic framework that we can always perform a canonical transformation of the form

$$\pi \mapsto \pi' := \pi + \Theta\,, \qquad (17.221)$$

where Θ is a closed 1-form on Riem(Σ). Closedness ensures that all Poisson brackets remain the same if π is replaced with π'. Since Riem(Σ) is an open positive convex cone in a vector space and hence contractible, it is immediate that $\Theta = \mathrm{d}f$ for some function $f : \mathrm{Riem}(\Sigma) \to \mathbb{R}$. However, π and π' must satisfy the diffeomorphism constraint, which is equivalent to saying that the kernel of π (considered as a 1-form on Riem(Σ)) contains the vector fields generated by spatial diffeomorphisms, which implies that Θ, too, must annihilate all those, so that f is constant on each connected component of the Diff(Σ) orbit in Riem(Σ). But, unless these orbits are connected, this does not imply that f is the pull back of a function on the quotient Riem(Σ)/ Diff(Σ), as assumed in [17.47]. We can only conclude that Θ is the pull back of a closed but not necessarily exact 1-form on superspace. Hence, there is an analog of the Bohm–Aharonov-like ambiguity that one always encounters if the configuration space is not simply connected. The quantum theory is then expected to display a sectorial structure labeled by the equivalence classes of unitary irreducible representations of the fundamental group of configuration space, which in analogy to Yang–Mills-type gauge theories are sometimes referred to as *θ-sectors* [17.50]. In GR the fundamental group of configuration space is isomorphic to a certain mapping-class group of the 3-manifold Σ. The theta structure then depends on the topology of Σ and can range from *trivial* to *very complicated*. See [17.51] for more details of the rôle and determination of these mapping-class groups and [17.39] for a more general discussion of the configuration space in GR, which, roughly speaking, is the quotient Riem(Σ)/ Diff(Σ), often referred to as *Wheeler's superspace* [17.35, 36, 52].

We finally note that additional theta structures may emerge if the gravitational field is formulated by means of different field variables including more mathematical degrees of freedom and more constraints (so as to result in the same number of physical degrees of freedom upon taking the quotient). The global structure of the additional gauge transformations may then add to the nontriviality of the fundamental group of configuration space and hence to the complexity of the sectorial structure. Examples have been discussed in the context of *Ashtekar variables* (Sect. 17.11) in connection with the *CP problem* in quantum gravity [17.53].

17.9 Asymptotic Flatness and Charges

Isolated systems are described by geometries which at large spatial distances approach a matter-free spacetime. In the case of vanishing cosmological constant the latter will be flat Minkowski spacetime. For nonzero Λ, it will be either de Sitter ($\Lambda > 0$) or anti-de Sitter ($\Lambda < 0$). Here we are interested in the case $\Lambda = 0$. We refer to Chap. 19) for the discussion of the anti-de Sitter case.

An initial data set (h, π) or (h, K) on Σ needs to satisfy certain asymptotic conditions in order to give rise to an asymptotically flat spacetime. Before going into this, we point out that there is also a topological condition on Σ in order to sensibly talk about *asymptotic regions*. The condition is that there exists a compact set $K \subset \Sigma$ such that its complement $\Sigma - K$ is diffeomorphic to the disjoint union of manifolds $\mathbb{R}^3 - B$, where B is a closed ball. These pieces into which Σ decomposes if one cuts out increasingly large compact sets are called *ends* of Σ. In passing, we note that the theory of *ends* for topological spaces and groups was developed by *Freudenthal* in 1931 [17.54]. Now, the first condition we pose is that there is only a finite number of such ends. (It is easy to see that manifolds may even have an uncountable number of ends.) With respect to each end, we can talk of approaching infinity. This means letting $r \to \infty$ if r is the stan-

dard radial coordinate on $\mathbb{R}^3 - B$ to which this end is diffeomorphic.

In order to make the Hamiltonian evolution generated by (17.205) well defined, we have to specify the boundary terms that we have to add in order to ensure functional differentiability with respect to h and π in the presence of long-ranging α and β. For the latter, we at least wish to include asymptotically constant α and covariant constant β corresponding to time and space translations, as well as asymptotic rotations.

For asymptotically constant α, we will pick up a boundary term from variations of (17.206a) with respect to h, which appears twice differentiated in the scalar curvature. The term itself is immediately read off from (17.105). In order to cancel it from the variation of H, where it appears multiplied with $\varepsilon(2\kappa)^{-1}$, we have to add the boundary term

$$-\varepsilon(2\kappa)^{-1}\alpha_\infty \int_{S^2_\infty} U^a\, m_a \sqrt{h}\,\mathrm{d}^2x\,, \tag{17.222}$$

where U^a is as in (17.105b) and α_∞ is the asymptotic value of α. The dominant contribution will come from the derivatives on h with all other factors of h being set to their asymptotic value $h_{ab} = \delta_{ab}$. Normalizing to $\alpha_\infty = 1$, this leads to the expression for the overall energy, called the ADM energy or ADM mass

$$\begin{aligned} E_{\text{ADM}} &= M_{\text{ADM}}c^2 \\ &= -\varepsilon(2\kappa)^{-1} \int_{S^2_\infty} (\partial_a h_{ab} - \partial_b h_{aa}) m_b\, \mathrm{d}\Omega\,, \end{aligned} \tag{17.223}$$

where all components refer to the asymptotically Euclidean coordinates for which indices are *raised* and *lowered* with δ_{ab}, so that we keep them on the same level, and where $\mathrm{d}\Omega$ is the rotationally invariant measure on S^2_∞, which in polar coordinates for the end is just $r^2 \sin\theta\, \mathrm{d}\theta \wedge \mathrm{d}\varphi$, and m_b are the components for the outward-pointing normal (normalized with respect to δ). Note that the dependence on the signature ε has to do with our earlier convention to keep the positive sign for the traceless modes in the kinetic-energy expression for both signatures.

The boundary terms for asymptotic β motions (translations and rotations) immediately follow from our earlier discussion: they are just the momentum maps for those motions, evaluated on the constraint surface. This gives

$$\begin{aligned} P_\beta|_{C_v=0} &= \left[\int_\Sigma \mathrm{d}^3x \pi^{ab} L_\beta h_{ab}\right]_{C_v=0} \\ &= 2\int_{S^2_\infty} \pi^{ab}\beta_a m_b\, \mathrm{d}\Omega\,. \end{aligned} \tag{17.224}$$

This leads to the linear momentum in the asymptotic β direction if β is taken to be an asymptotically covariant constant and normalized vector field, like $\beta_a = \delta_{ab}$ for the translation in the b direction, and to angular momentum in the ω direction if $\beta_a = \varepsilon_{abc}\omega_b x_c$, with ω normalized.

For these definitions to make sense, we have to clarify the issues of existence and uniqueness. Existence means that we have to make sure that these integrals exist. Since an integral over S^2_∞ is the limit of $S^2(r)$ integrals in the limit $r \to \infty$, this means proving that the limit exists. Looking at (17.223) and (17.224), we see that we need a r^{-2} fall off for the combination $(\partial_b h_{ab} - \partial_a h_{bb}) m_a$ and likewise for $\pi^{ab}\beta_a m_b$. The question is what conditions this implies for the fields h and π, given that they are solutions to the constraints. Moreover, given existence for certain fall-off conditions, we also want them to ensure uniqueness, meaning that the calculated asymptotic charges are the same for any two different asymptotically Euclidean coordinate systems in which the fall-off conditions hold. Finally, we want these quantities to be preserved under the Hamiltonian evolution, i. e. to be conserved charges. The main result in this direction, at least as far as energy and linear momentum are concerned (i. e. ignoring angular momentum for the moment), is the following: ADM energy and linear momentum exist uniquely and are preserved under Hamiltonian evolution if

$$h_{ab} = \delta_{ab} + o_2(r^{-1/2})\,, \tag{17.225a}$$

$$\pi^{ab} = o_1(r^{-3/2})\,. \tag{17.225b}$$

Here we employ the *little-o notation*, where $o_p(r^{-\alpha})$ stands for terms with fall off faster than $r^{-\alpha}$ and whose q-th derivatives fall off faster than $r^{-\alpha-q}$ for all $q \le p$. These are also the conditions for which stability of Minkowski space is known to hold [17.55]. See Chap. 18 for an example showing that the fall off for h faster than $r^{-1/2}$ cannot be further relaxed.

At first sight (17.225a) might seem too weak to guarantee existence of (17.223). The reason why it

is not is, in fact, easy to see. If we convert (17.223) into a bulk integral using Gauss' theorem, the integrand contains a combination of second derivatives of h which just form the second derivative part of the scalar curvature. This is how we derived the expression (17.223) from (17.105) in the first place. Using the scalar constraint, we can express this combination by terms quadratic in the first derivatives of h and quadratic in π, whose bulk integrals exist due to (17.225). See also [17.56] for a more comprehensive discussion.

For asymptotically flat and stationary solutions, the ADM mass M_{ADM} is known to coincide with the so-called Komar mass [17.57], whose simple and coordinate-invariant expression is

$$M_{\text{Komar}} = \frac{-\varepsilon}{\kappa c^2} \int_{S^2_\infty} \star \, \mathrm{d}K^\flat \, . \tag{17.226}$$

There exist various proofs in the literature showing $M_{\text{Komar}} = M_{\text{ADM}}$; see, e.g., [17.58–60] and [17.61, Theorem 4.13].

Last but not least, we mention the positive-mass theorem (for Lorentzian signature $\varepsilon = -1$), which states that for any pair (h, π) of initial data satisfying the constraints, $M_{\text{ADM}} \geq 0$, with equality only if the data are that of Minkowski space. Note that the expression (17.223) for M_{ADM} is a functional of h alone, but that in the formulation of the positive-mass theorem given here it is crucial that for h there exists a π so that the pair (h, π) solves the constraint. Otherwise it is easy to write down 3-metrics with negative ADM mass; take e.g. (17.232) (see below) for negative r_0, suitably smoothed out for smaller radii so as to avoid the singularity at $r = -r_0$. Since the ADM mass only depends on the asymptotic behavior, it is completely independent of any alterations to the metric in the interior. If one wishes to make the positive-mass theorem a statement about metrics alone without any reference to the constraints, one has to impose positivity conditions on the scalar curvature. But that also imposes topological restrictions due to the result of *Gromov* and *Lawson* [17.32] mentioned at the end of Sect. 17.6. The positive-mass theorem is discussed in detail in Chap. 18.

In passing, we remark that the positive-mass theorem in combination with the equality $M_{\text{ADM}} = M_{\text{Komar}}$ gives a simple proof of the absence of *gravitational solitons*, i.e. stationary asymptotically flat solutions to Einstein's equations on $\Sigma = \mathbb{R}^3$. This follows from (17.226) and $\mathrm{d} \star \mathrm{d}K^\flat \propto i_k\mathbf{Ric}$. The vacuum equation $\mathbf{Ric} = 0$ then implies that $M_{\text{ADM}} = M_{\text{Komar}} = 0$, which implies that spacetime is flat Minkowski. This theorem was originally shown for static spacetimes (i.e. hypersurface orthogonal K) by *Einstein* and *Pauli* [17.62] and later generalized to the stationary case by *Lichnerowicz* [17.63]. The result of this theorem cannot be circumvented by trying more complicated topologies for Σ. As soon as Σ becomes nonsimply connected (which, in view of the validity of the Poincaré conjecture, will be the case for any one-ended manifold other than $\mathbb{R}^3$), we know from *Gannon*'s theorem [17.64] that the evolving spacetime will inevitably develop singularities.

Finally, we mention that under suitable fall-off conditions we can find the Poincaré group as an *asymptotic symmetry group* [17.65]. It will emerge from (17.219) as equivalence classes of all hypersurface deformations, including those in which α and β asymptotically approach rigid translations, rotations, or boosts. The quotient is taken with respect to those deformations which are generated by the constraints, in which α and β tend to zero at spatial infinity. There are various subtleties and fine tunings involved for the precise fall-off conditions that are necessary in order to exactly obtain a 10-dimensional symmetry as a quotient of two infinite-dimensional objects. This is particularly true for asymptotic boosts, for which one needs to tilt the hypersurface, corresponding to asymptotic lapse functions $\alpha \propto r$. (Boosted hypersurfaces are known to exist in the development of asymptotically flat initial data [17.66].) But, leaving the analytic details aside, the qualitative picture is quite generic for gauge field theories with long-ranging field configurations [17.67]: a proper physical symmetry group arises as a quotient of a general covariance group with respect to a proper normal subgroup, the latter being defined to be that object that is generated by the constraints.

17.10 Black-Hole Data

In this section, we discuss some simple solutions to the vacuum Einstein equations without cosmological constant. We first specify to the simplest case of time-symmetric conformally flat data. Time symmetry means that the initial extrinsic curvature vanishes, $K = 0$. The corresponding Cauchy surface will then be totally

geodesic in the spacetime that emerges from it. The vector constraint (17.131a) is identically satisfied and the scalar constraint (17.131b) reduces to scalar flatness

$$R(h) = \mathbf{Scal}^D = 0\,. \tag{17.227}$$

Conformal flatness means that

$$h = \Omega^4 \delta\,, \tag{17.228}$$

where δ is the flat metric. From (17.112a), we infer that (17.227) is equivalent to Ω being harmonic

$$\Delta_\delta \Omega = 0\,, \tag{17.229}$$

where Δ_δ is the Laplacian with respect to the flat metric δ. We seek solutions Ω which are asymptotically flat for $r \to \infty$ and give rise to complete manifolds in the metric structure defined by g. The only spherically symmetric such solution is

$$\Omega(r) = 1 + \frac{r_0}{r}\,, \tag{17.230}$$

where the integration constant r_0 can be related to the ADM mass (17.223) by

$$M_{\mathrm{ADM}} = 2c^2 r_0/G\,. \tag{17.231}$$

This solution is defined on $\Sigma = \mathbb{R}^3 - \{0\}$. The metric on Σ so obtained is

$$h = \left(1 + \frac{r_0}{r}\right)^4 \left(\mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2)\right)\,. \tag{17.232}$$

It admits the following isometries

$$I_1(r,\theta,\varphi) := \left(r_0^2/r, \theta, \varphi\right)\,, \tag{17.233a}$$

$$I_2(r,\theta,\varphi) := \left(r_0^2/r, \pi - \theta, \varphi + \pi\right)\,. \tag{17.233b}$$

Note that the second is just a composition of the first with the antipodal map $(r,\theta,\varphi) \mapsto (r, \pi - \theta, \varphi + \pi)$, which is well defined on $\mathbb{R}^3 - \{0\}$. This makes I_2 a fixed-point free action. The fixed-point set of I_1 is the 2-sphere $r = r_0$. Note that generally a submanifold that is the fixed-point set of an isometry is necessarily totally geodesic (has vanishing extrinsic curvature). To see this, consider a geodesic that starts on and tangentially to this submanifold. Such a geodesic cannot leave the submanifold, for if it did we could use the isometry to map it to a different geodesic with identical initial conditions, in contradiction to the uniqueness of solutions for the geodesic equation. Hence, the 2-sphere $r = r_0$ has vanishing extrinsic curvature and is, therefore, in particular, a minimal surface (has vanishing trace of the extrinsic curvature). The geometry inside the sphere $r = r_0$ is isometric to that outside it. This is depicted in Fig. 17.4.

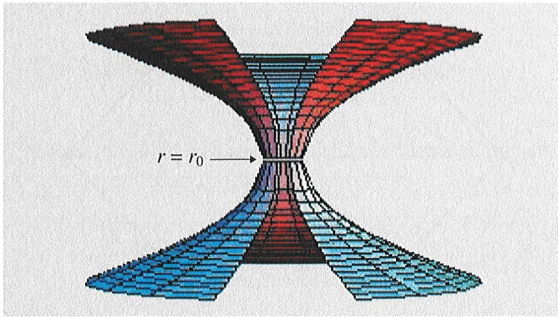


Fig. 17.4 Cauchy surface with time-symmetric initial data and two isometric asymptotically flat ends separated by a totally geodesic 2-sphere

For the data ($h =$ (17.232), $K = 0$) on $\Sigma = \mathbb{R}^3 - \{0\}$, we actually know its maximal time evolution: it is the Kruskal spacetime [17.68, 69], which maximally extends the exterior Schwarzschild spacetime. Figure 17.5 shows a conformal diagram of Kruskal spacetime.

In Kruskal coordinates (*Kruskal* [17.68] uses (v,u), *Hawking* and *Ellis* [17.69] (t',x') for what we call (T,X)) (T,X,θ,φ), where T and X each range in $(-\infty,\infty)$ obeying $T^2 - X^2 < 1$, the Kruskal metric reads (as usual, we write $\mathrm{d}\Omega^2$ for $\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2$)

$$g = \frac{8r_0^2}{r}\exp(-r/r_0)\left(-\mathrm{d}T^2 + \mathrm{d}X^2\right) + r^2\mathrm{d}\Omega^2\,, \tag{17.234}$$

where r is a function of T and X, implicitly defined by

$$((r/r_0) - 1)\exp(r/r_0) = X^2 - T^2\,. \tag{17.235}$$

The metric is spherically symmetric and allows for the additional Killing field

$$K = (X\partial_T + T\partial_X)\,, \tag{17.236}$$

which is timelike for $|X| > |T|$ and spacelike for $|X| < |T|$.

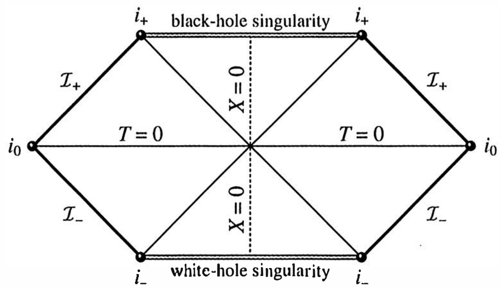


Fig. 17.5 Conformally compactified Kruskal spacetime. The T axis points up vertically, the X axis horizontally to the right. The Cauchy surface of Fig. 17.4 corresponds to the hypersurface $T=0$. The various infinities are: i_0 spacelike, $i_\pm$ future/past timelike, and $\mathcal{I}_\pm$ future/past lightlike infinity. The *diamond-shaped* region at the right of the figure corresponds to the usual exterior Schwarzschild solution containing one asymptotically flat end

Both maps (17.233) extend to the Kruskal manifold. The fixed-point free action (17.233b) has the extension

$$J:(T,X,\theta,\varphi)\mapsto(T,-X,\pi-\theta,\varphi+\pi)\,. \quad (17.237)$$

It generates a freely acting group $\mathbb{Z}_2$ of smooth isometries which preserve space as well as time orientation. Hence, the quotient is a smooth space- and time-orientable manifold, that is sometimes called the $\mathbb{R}P^3$-*geon*. It represents the maximal time evolution of the data ($h=(17.232)$, $K=0$) as above, but now defined on the initial quotient manifold $\Sigma=(\mathbb{R}^3-\{0\})/I_2$. It has only one asymptotically flat end and the topology of a once-punctured real projective space $\mathbb{R}P^3$. Note that the map J preserves the Killing field (17.236) only up to sign. Had one chosen $J':(T,X,\theta,\varphi)\mapsto(-T,-X,\pi-\theta,\varphi+\pi)$ as in [17.70, 71], one would have preserved K but lost time orientability.

Within the set of conformally flat and time-symmetric initial data we can easily generalize the solution (17.230) to (17.229) to include more than one monopole term on a multipunctured $\mathbb{R}^3$. For two terms, we get

$$\Omega(r)=1+\frac{a_1}{r_1}+\frac{a_2}{r_2}\,, \quad (17.238)$$

where $r_i=\|\vec{x}-\vec{c}_i\|$. This represents two black holes without spin and orbital angular momentum momentarily at rest. The manifold has three ends, one for $r\to\infty$ and one each for $r_i\to 0$. For each end we can calculate the ADM mass and get

$$M=2(a_1+a_2)c^2/G\,, \quad (17.239a)$$

$$M_1=2\left(a_1+\frac{a_1a_2}{r_{12}}\right)\frac{c^2}{G}\,, \quad (17.239b)$$

$$M_2=2\left(a_2+\frac{a_1a_2}{r_{12}}\right)\frac{c^2}{G}\,, \quad (17.239c)$$

where $r_{12}:=\|\vec{c}_1-\vec{c}_2\|$. Here M is the total mass associated with the end $r\to\infty$ and M_i is the individual hole mass associated with the end $r_i\to 0$. The binding energy is the overall energy minus the individual ones. One obtains

$$\Delta E:=(M-M_1-M_2)c^2=-G\frac{M_1M_2}{r_{12}}+\dots\,, \quad (17.240)$$

where the dots stand for corrections of quadratic and higher powers in GM_i/c^2r_{12}. This can be easily generalized to any finite number of poles. Note that the initial manifolds are all complete, i. e. all punctures lie at infinite metric distance from any interior point.

Other generalizations consist in adding linear and angular momenta. This can be done using the conformal method, which we briefly describe. We maintain conformal flatness (17.228) and set for the extrinsic curvature

$$K_{ab}=\Omega^{-2}\bar{K}_{ab}+\tfrac{1}{3}\Omega^4\delta_{ab}\tau\,, \quad (17.241)$$

where τ is constant. The vector constraint in vacuum and for $\lambda=0$ is then satisfied if $\bar{K}$ is transverse and traceless with respect to the flat metric δ

$$\delta_{ab}\bar{K}_{ab}=0\,, \quad (17.242a)$$

$$\partial_a\bar{K}_{ab}=0\,. \quad (17.242b)$$

Once such a $\bar{K}$ is found, the scalar constraint determines the conformal factor Ω via

$$-\varepsilon\Delta_\delta\Omega+\tfrac{1}{8}\Omega^{-7}\bar{K}_{ab}\bar{K}_{ab}-\tfrac{1}{12}\Omega^5\tau^2=0\,, \quad (17.243)$$

which follows from the scalar constraint once the ansatz (17.241) is inserted and the conformal flatness of h is used to express the scalar curvature according to (17.112). Existence and uniqueness of this equation

for the Lorentzian case $\varepsilon = -1$ will be discussed in Chap. 16.

It is remarkable that the ADM (17.224) can be calculated without knowing Ω. Hence, we can parameterize solutions to (17.242) directly by the momenta without solving (17.243) first. Two solutions are the *Bowen–York data* [17.34, 72]

$$\bar{K}^{(1)}_{ab} = r^{-2}\left(\nu_a A_b + \nu_b A_a - (\delta_{ab} - \nu_a \nu_b)\nu_c A_c\right), \tag{17.244}$$

$$\bar{K}^{(2)}_{ab} = r^{-3}\left(\nu_a \varepsilon_{bcd} + \nu_b \varepsilon_{acd}\right) B_c \nu_d, \tag{17.245}$$

where $\nu_a := x_a/r$ and A and B are constant vectors. One verifies directly that these $\bar{K}$ are transverse traceless. Using (17.224), one shows that (17.244) has vanishing angular momentum and a linear momentum with components

$$P_a = \frac{2c^3}{3G} A_a, \tag{17.246}$$

whereas (17.245) has vanishing linear momentum and an angular momentum with components

$$J_a = \frac{c^3}{3G} B_a. \tag{17.247}$$

They can be combined to give data for single holes with nonzero linear and angular momenta and also be superposed in order to give data for multi-black-hole configurations. Such data, and certain modifications of them, form the essential ingredient for present-day numerical simulations of black-hole scattering and the subsequent emission of gravitational radiation.

17.11 Further Developments, Problems, and Outlook

In this contribution we have explained in some detail the dynamical and Hamiltonian formulation of GR. We followed the traditional ADM approach in which the basic variables are the Riemannian metric h of space and its conjugate momentum π, which is essentially the extrinsic curvature that Σ will assume once the spacetime is developed and Σ is isometrically embedded in it. Attempts to establish a theory of *quantum gravity* based on the Hamiltonian formulation of GR suggest that other canonical variables are better suited for the mathematical implementation of the constraints and the ensuing construction of spaces of states and observables [17.23, 73, 74]. These variables are a (suitably densitized) orthonormal 3-bein field E on Σ and the *Ashtekar–Barbero connection.* We have already seen that orientable Σ are parallelizable so that global fields E do indeed exist. Any field E determines a Riemannian metric h, which in turn determines its Levi-Civita connection. The Ashtekar–Barbero covariant derivative, $\mathcal{D}$, differs from the Levi-Civita connection D of h by the endomorphism-valued 1-form which associates to each tangent vector X the tangent-space endomorphism $Y \mapsto \gamma\mathbf{Wein}(X) \times Y$, where γ is a dimensionless constant, the so-called Barbero–Immirzi parameter. Hence, we have

$$\mathcal{D}_X Y = D_X Y + \gamma \mathbf{Wein}(X) \times Y. \tag{17.248}$$

The multiplication $\times$ is the standard three-dimensional vector product with respect to the metric h. It is defined as follows

$$X \times Y := [\star(X^\flat \wedge Y^\flat)]^\sharp, \tag{17.249}$$

where the isomorphisms $\flat$ and $\sharp$ are with respect to h (cf. (17.1)). The product $\times$ obeys the standard rules: it is bilinear, anti-symmetric, and $X \times (Y \times Z) = h(X,Z)Y - h(X,Y)Z$. Moreover, for any X, the endomorphism $Y \mapsto X \times Y$ is anti-symmetric with respect to h, i. e. $h(X \times Y, Z) = -h(Y, X \times Z)$, and hence it is in the Lie algebra of the orthogonal group of h. In particular, this is true for $Y \mapsto \mathbf{Wein}(X) \times Y$, showing that $\mathcal{D}$ is again metric, i. e. obeys $\mathcal{D}h = 0$ once its unique extension to all tensor fields is understood. Clearly, unlike D, the torsion of $\mathcal{D}$ cannot be zero

$$\begin{aligned} T^{\mathcal{D}}(X,Y) &= \mathcal{D}_X Y - \mathcal{D}_Y X - [X,Y] \\ &= \gamma\left(\mathbf{Wein}(X) \times Y - \mathbf{Wein}(Y) \times X\right). \end{aligned} \tag{17.250}$$

Using (17.66) and index notation, the curvature tensor for $\mathcal{D}$ is

$$\begin{aligned} R^{\mathcal{D}}_{abcd} = {} & R^{D}_{abcd} \\ & + \varepsilon\gamma\left(D_c K_{dn} - D_d K_{cn}\right)\varepsilon^{n}{}_{ab} \\ & - \gamma^2\left(K_{ac}K_{bd} - K_{ad}K_{bc}\right). \end{aligned} \tag{17.251}$$

From this, the scalar curvature follows

$$\mathbf{Scal}^{\mathcal{D}} = \mathbf{Scal}^{D} + \gamma^2 G^{abcd} K_{ab} K_{cd}. \tag{17.252}$$

Comparison with (17.131a) shows that for $\gamma^2 = \varepsilon$ the gravitational part of the scalar constraint is just (ε times) the scalar curvature of $\mathcal{D}$. This striking simplification of the scalar constraint formed the original motivation for the introduction of $\mathcal{D}$ by *Ashtekar* [17.75]. However, for $\varepsilon = -1$ one needs to complexify the tensor bundle over Σ for $\gamma = \pm i$ to make sense, and subsequently impose reality conditions which re-introduce a certain degree of complication; see, e.g., [17.76] for a compact account not using spinors. The usage of $\mathcal{D}$ in the real case was then proposed by *Barbero* in [17.77] and forms the basic tool in *loop quantum gravity* [17.73], which has definite technical advantages over the metric-based traditional approach.

On the other hand, the traditional approach is well suited to address certain conceptual problems [17.37], like e.g. the problem of time that emerges in those cases where the Hamiltonian (17.205) has no boundary terms and is therefore just a sum of constraints. This happens in cosmology based on closed Σ. The motions generated by the Hamiltonian are then just pure gauge transformations and the question arises of whether and how *motion* and *change* are to be recovered (compare Chap. 36). Dynamical models in cosmology often start from symmetry assumptions that initially reduce the infinitely many degrees of freedom to finitely many ones (so-called mini-superspace models). Other modes are then treated perturbatively in an expansion around the symmetric configurations. In these cases quantization in the metric representation can be performed, with potentially interesting consequences for observational cosmology, like the modification of the anisotropy spectrum of the cosmic microwave background [17.78, 79]. All these attempts make essential use of the Hamiltonian theory as described in this contribution.

References

17.1 P.A.M. Dirac: The theory of gravitation in Hamiltonian form, Proc. R. Soc. A **246**(1246), 333–343 (1958)

17.2 P.A.M. Dirac: Generalized Hamiltonian dynamics, Proc. R. Soc. A **246**(1246), 326–332 (1958)

17.3 P.A.M. Dirac: *Lectures on Quantum Mechanics*, Belfer Graduate School of Science Monographs, Vol. 2 (Yeshiva Univ., New York 1964)

17.4 M. Gotay, J. Nester, G. Hinds: Presymplectic manifolds and the Dirac–Bergmann theory of constraints, J. Math. Phys. **19**(11), 2388–2399 (1978)

17.5 M. Henneaux, C. Teitelboim: *Quantization of Gauge Systems* (Princeton Univ. Press, Princeton 1992)

17.6 R. Arnowitt, S. Deser: Quantum theory of gravitation: General formulation and linearized theory, Phys. Rev. **113**(2), 745–750 (1959)

17.7 R. Arnowitt, S. Deser, C.W. Misner: Dynamical structure and definition of energy in general relativity, Phys. Rev. **116**(5), 1322–1330 (1959)

17.8 R. Arnowitt, S. Deser, C.W. Misner: Canonical variables for general relativity, Phys. Rev. **117**(6), 1595–1602 (1960)

17.9 R. Arnowitt, S. Deser, C.W. Misner: Canonical variables, expressions for energy, and the criteria for radiation in general relativity, Nuovo Cim. **15**(3), 487–491 (1960)

17.10 R. Arnowitt, S. Deser, C.W. Misner: Finite self-energy of classical point particles, Phys. Rev. Lett. **4**(7), 375–377 (1960)

17.11 R. Arnowitt, S. Deser, C.W. Misner: Energy and the criteria for radiation in general relativity, Phys. Rev. **118**(4), 1100–1104 (1960)

17.12 R. Arnowitt, S. Deser, C.W. Misner: Consistency of the canonical reduction of general relativity, J. Math. Phys. **1**(5), 434–439 (1960)

17.13 R. Arnowitt, S. Deser, C.W. Misner: Gravitational-electromagnetic coupling and the classical self-energy problem, Phys. Rev. **120**(1), 313–320 (1960)

17.14 R. Arnowitt, S. Deser, C.W. Misner: Interior Schwarzschild solutions and interpretation of source terms, Phys. Rev. **120**(1), 321–324 (1960)

17.15 R. Arnowitt, S. Deser, C.W. Misner: Note on positive-definiteness of the energy of the gravitational field, Ann. Phys. **11**(1), 116–121 (1960)

17.16 R. Arnowitt, S. Deser, C.W. Misner: Heisenberg representation in classical general relativity, Nuovo Cim. **19**(4), 668–681 (1961)

17.17 R. Arnowitt, S. Deser, C.W. Misner: Wave zone in general relativity, Phys. Rev. **121**(5), 1556–1566 (1961)

17.18 R. Arnowitt, S. Deser, C.W. Misner: Coordinate invariance and energy expressions in general relativity, Phys. Rev. **122**(3), 997–1006 (1961)

17.19 R. Arnowitt, S. Deser, C.W. Misner: The dynamics of general relativity. In: *Gravitation: An Introduction to Current Research*, ed. by L. Witten (Wiley, New York, London 1962) pp. 227–265, arXiv:gr-qc/0405109

17.20 R. Arnowitt, S. Deser, C.W. Misner: Republication of: The Dynamics of general relativity, Gen. Relativ. Gravit. **40**(9), 1997–2027 (2008), Republication as *Golden Oldie* with some minor corrections. Available online as arXiv:gr-qc/0405109

17.21 J. Pullin: Editorial note to R. Arnowitt, S. Deser, C.W. Misner: The dynamics of general relativity, Gen. Relativ. Gravit. **40**(9), 1989–1995 (2008)

17.22 É. Gourgoulhon: *3+1 Formalism in General Relativity*, Lecture Notes in Physics, Vol. 846 (Springer, Berlin 2012)

17.23 M. Bojowald: *Canonical Gravity and Applications. Cosmology, Black Holes, Quantum Gravity* (Cambridge Univ. Press, Cambridge 2011)
17.24 R. Geroch: Domain of dependence, J. Math. Phys. **11**(2), 437–449 (1970)
17.25 N. Steenrod: *The Topology of Fibre Bundles* (Princeton Univ. Press, Princeton 1951)
17.26 J.W. Milnor, J.W. Stasheff: *Characteristic Classes*, Annals of Mathematics Studies, Vol. 76 (Princeton Univ. Press, Princeton 1974)
17.27 J.H.C. Whitehead: The immersion of an open 3-manifold in Euclidean 3-space, Proc. Lond. Math. Soc. (3) **11**(1), 81–90 (1961)
17.28 R. Geroch: Spinor structure of spacetimes in general relativity I, J. Math. Phys. **9**(11), 1739–1744 (1968)
17.29 M. Spivak: *A Comprehensive Introduction to Differential Geometry I–V* (Publish or Perish, Wilmington, Delaware 1979)
17.30 J.L. Kazdan, F.W. Warner: Scalar curvature and conformal deformation of Riemannian structure, J. Differ. Geom. **10**(1), 113–134 (1975)
17.31 D. Witt: Vacuum space-times that admit no maximal slices, Phys. Rev. Lett. **57**(12), 1386–1389 (1986)
17.32 M. Gromov, B. Lawson: Positive scalar curvature and the Dirac operator on complete Riemannian manifolds, Math. Inst. Ht. Études Sci. **58**(1), 83–196 (1983)
17.33 D. Giulini: 3-Manifolds for relativists, Int. J. Theor. Phys. **33**, 913–930 (1994)
17.34 J.W. York: Kinematics and dynamics of general relativity. In: *Sources of Gravitational Radiation*, ed. by L. Smarr (Cambridge Univ. Press, Cambridge 1979) pp. 83–126
17.35 B.S. DeWitt: Quantum theory of gravity. I. The canonical theory, Phys. Rev. **160**(5), 1113–1148 (1967)
17.36 B.S. DeWitt: Erratum, Phys. Rev. **171**(5), 1834 (1968)
17.37 C. Kiefer: *Quantum Gravity*, International Series of Monographs on Physics, Vol. 124, 2nd edn. (Clarendon, Oxford 2007)
17.38 D. Giulini: What is the geometry of superspace? Phys. Rev, D **51**(10), 5630–5635 (1995)
17.39 D. Giulini: The superspace of geometrodynamics, Gen. Relativ. Gravit. **41**(4), 785–815 (2009)
17.40 C. Blohmann, M.C.B. Fernandes, A. Weinstein: Groupoid symmetry and constraints in general relativity, Commun. Contemp. Math. **15**(01), 1250061 (2013)
17.41 T. Regge, C. Teitelboim: Role of surface integrals in the Hamiltonian formulation of general relativity, Ann. Phys. **88**, 286–318 (1974)
17.42 F.G. Markopoulou: Gravitational constraint combinations generate a Lie algebra, Class. Quantum Gravity **13**(9), 2577–2584 (1996)
17.43 C. Teitelboim: How commutators of constraints reflect the spacetime structure, Ann. Phys. **79**(2), 542–557 (1973)
17.44 S.A. Hojman, K. Kuchař, C. Teitelboim: Geometrodynamics regained, Ann. Phys. **96**, 88–135 (1976)
17.45 C.J. Isham, K.V. Kuchař: Representations of spacetime diffeomorphisms. I. Canonical parametrized field theories, Ann. Phys. **164**, 288–315 (1985)
17.46 C.J. Isham, K.V. Kuchař: Representations of spacetime diffeomorphisms. II. Canonical geometrodynamics, Ann. Phys. **164**, 316–333 (1985)
17.47 K. Kuchař: Geometrodynamics regained: A Lagrangian approach, J. Math. Phys. **15**(6), 708–715 (1974)
17.48 D. Lovelock: The four-dimensionality of space and the Einstein tensor, J. Math. Phys. **13**(6), 874–876 (1972)
17.49 P. Hořava: Quantum gravity at a Lifshitz point, Phys. Rev. **79**(8), 084008 (2009)
17.50 C.J. Isham: *Theta*-states induced by the diffeomorphism group in canonically quantized gravity, Quantum Structure of Space And Time. Proc. Nuffield Workshop, Imp. College Lond., ed. by J.J. Duff, C.J. Isham (Cambridge Univ. Press, London 1982) pp. 37–52
17.51 D. Giulini: Mapping-class groups of 3-manifolds in canonical quantum gravity. In: *Quantum Gravity: Mathematical Models and Experimental Bounds*, ed. by B. Fauser, J. Tolksdorf, E. Zeidler (Birkhäuser, Basel 2007), available online at arxiv.org/pdf/gr-qc/0606066
17.52 J.A. Wheeler: Geometrodynamics and the issue of the final state. In: *Relativity, Groups and Topology. 1963 Les Houches Lectures*, ed. by C.M. DeWitt, B.S. DeWitt (Gordon and Breach, New York 1964) pp. 317–520
17.53 A. Ashtekar, A.P. Balachandran, S.G. Jo: The CP problem in quantum gravity, Int. J. Mod. Phys. A **4**(6), 1493–1514 (1989)
17.54 H. Freudenthal: Über die Enden topologischer Räume und Gruppen, Math. Ann. **33**(1), 692–713 (1931)
17.55 L. Bieri, N. Zipser: *Extensions of the Stability Theorem of the Minkowski Space in general relativity*, Studies in Advanced Mathematics, Vol. 45 (American Mathematical Society, Providence 2009)
17.56 N.Ó. Murchadha: Total energy momentum in general relativity, J. Math. Phys. **27**(8), 2111–2128 (1986)
17.57 A. Komar: Covariant conservation laws in general relativity, Phys. Rev. **113**(3), 934–936 (1959)
17.58 R. Beig: Arnowitt–Deser–Misner energy and g_{00}, Phys. Lett. A **69**(3), 153–155 (1978)
17.59 A. Ashtekar, A. Magnon-Ashtekar: On conserved quantities in general relativity, J. Math. Phys. **20**(5), 793–800 (1979)
17.60 P. Chruściel: A remark on the positive-energy theorem, Class. Quantum Gravity **3**(6), L115–L121 (1986)
17.61 Y. Choquet-Bruhat: *General Relativity and the Einstein Equations* (Oxford Univ. Press, Oxford 2009)
17.62 A. Einstein, W. Pauli: On the non-existence of regular stationary solutions of relativistic field equations, Ann. Math. **44**(2), 131–137 (1943)

17.63 A. Lichnerowicz: *Théories Relativistes de la Gravitation et de l'Électromagnétisme* (Masson et Cie, Paris 1955)

17.64 D. Gannon: Singularities in nonsimply connected space-times, J. Math. Phys. **16**(12), 2364–2367 (1975)

17.65 R. Beig, N.Ó. Murchadha: The Poincaré group as symmetry group of canonical general relativity, Ann. Phys. **174**, 463–498 (1987)

17.66 D. Christodoulou, N.Ó. Murchadha: The boost problem in general relativity, Commun. Math. Phys. **80**(2), 271–300 (1981)

17.67 D. Giulini: Asymptotic symmetry groups of long-ranged gauge configurations, Mod. Phys. Lett. A **10**(28), 2059–2070 (1995)

17.68 M.D. Kruskal: Maximal extension of Schwarzschild metric, Phys. Rev. **119**(5), 1743–1745 (1960)

17.69 S.W. Hawking, G.F.R. Ellis: *The Large Scale Structure of Spacetime* (Cambridge Univ. Press, Cambridge 1973)

17.70 C. Misner, J.A. Wheeler: Classical physics as geometry: Gravitation, electromagnetism, unquantized charge, and mass as properties of curved empty space, Ann. Phys. **2**, 525–660 (1957)

17.71 G.W. Gibbons: The elliptic interpretation of black holes and quantum mechanics, Nucl. Phys. B **98**, 497–508 (1986)

17.72 J.M. Bowen, J.W. York Jr.: Time-asymmetric initial data for black holes and black-hole collisions, Phys. Rev. D **21**(8), 2047–2056 (1980)

17.73 T. Thiemann: *Modern Canonical Quantum General Relativity*, Cambridge Monographs on Mathematical Physics (Cambridge Univ. Press, Cambridge 2007)

17.74 C. Rovelli: *Quantum Gravity*, Cambridge Monographs on Mathematical Physics (Cambridge Univ. Press, Cambridge 2004)

17.75 A. Ashtekar: New Hamiltonian formulation of general relativity, Phys. Rev. D **36**(6), 1587–1602 (1987)

17.76 D. Giulini: Ashtekar variables in classical general relativity. In: *Canonical Gravity: From Classical to Quantum*, Lecture Notes in Physics, Vol. 434, ed. by J. Ehlers, H. Friedrich (Springer, Berlin, 1994) pp. 81–112

17.77 F. Barbero: Real Ashtekar variables for Lorentzian signature space-times, Phys. Rev. D **51**(10), 5507–5510 (1995)

17.78 C. Kiefer, M. Krämer: Quantum gravitational contributions to the cosmic microwave background anisotropy spectrum, Phys. Rev. Lett. **108**(2), 021301 (2012)

17.79 D. Bini, G. Esposito, C. Kiefer, M. Krämer, F. Pessina: On the modification of the cosmic microwave background anisotropy spectrum from canonical quantum gravity, Phys. Rev. D **87**(10), 104008 (2013)

18. Positive Energy Theorems in General Relativity

Sergio Dain

At the end of the nineteenth century light was regarded as an electromagnetic wave propagating in a material medium called *ether*. The speed c appearing in Maxwell's wave equations was the speed of light with respect to the ether. Therefore, according to the Galilean addition of velocities, the speed of light in the laboratory would differ from c. The measure of such a difference would reveal the motion of the laboratory (the Earth) relative to the ether (a sort of *absolute motion*). However, the Earth's absolute motion was never evidenced.

Galileo addition of velocities is based on the assumption that lengths and time intervals are *invariant* (independent of the state of motion). In this way of thinking, the spacetime emanates from our daily experience and lies at the heart of Newton's classical mechanics. Nevertheless, in 1905 Einstein defied Galileo addition of velocities by postulating that light travels at the same speed c in any inertial frame. In doing so, Einstein extended the *principle of relativity* to the electromagnetic phenomena described by Maxwell's laws. In Einstein's special relativity the ether does not exist and the absolute motion is devoid of meaning. The invariance of the speed of light forced the replacement of Galileo transformations with Lorentz transformations. Thus, relativistic length contractions and time dilations entered our understanding of spacetime. Newtonian mechanics had to be reformulated, which led to the discovery of the mass-energy equivalence.

18.1 Theorems

The aim of this chapter is to present an introduction and also an overview of some of the most relevant results concerning positivity energy theorems in general relativity. These theorems provide the answer to a long standing problem that has been proved remarkably difficult to solve. They constitute one of the major results in classical general relativity and they uncover a deep self-consistence of the theory.

In this introductory section we would like to present the theorems in a complete form but with the least possible amount of technical details, in such a way that the reader can have a rough idea of the basic ingredients. The examples that illustrate the hypothesis of the theorems are discussed in the following sections.

An isolated system is an idealization in physics that assumes that the sources are confined to a finite region and the fields are weak far away from the sources. This kind of system is expected to have finite total energy. In general relativity there are several ways of defining isolated systems. For our purpose the most appropriate definition is through initial conditions for Einstein equations. The reasons for this are twofold. First, the notion of total energy has been discovered and formulated using a Hamiltonian formulation of the theory which involves the study of initial conditions. We refer the reader to Chap. 17 for this topic. Second, the proofs of the positive mass theorem are mainly given in terms of initial conditions. For a discussion of the ini-

tial value formulation of Einstein equations we refer to Chap. 16.

Initial conditions for Einstein equations are characterized by an *initial data set* given by $(S, h_{ij}, K_{ij}, \mu, j^i)$, where S is a connected three-dimensional manifold, h_{ij} a (positive definite) Riemannian metric, K_{ij} a symmetric tensor field, μ a scalar field, and j^i a vector field on S, such that the constraint equations

$$D_j K^{ij} - D^i K = -8\pi j^i\,, \tag{18.1}$$

$$R - K_{ij}K^{ij} + K^2 = 16\pi\mu\,, \tag{18.2}$$

are satisfied on S. Here D and R are the Levi-Civita connection and scalar curvature associated with h_{ij}, and $K = K_{ij}h^{ij}$. In these equations the indices $i, k, \ldots$ are three-dimensional indices; they are raised and lowered with the metric h_{ij} and its inverse h^{ij}. The matter fields are assumed to satisfy the dominant energy condition

$$\mu \geq \sqrt{j^i j_i}\,. \tag{18.3}$$

The initial data model an isolated system if the fields are weak far away from sources. This physical idea is captured in the following definition of an asymptotically flat initial data set. Let B_R be a ball of finite radius R in $\mathbb{R}^3$. The exterior region $U = \mathbb{R}^3 \setminus B_R$ is called an *end*. On U we consider Cartesian coordinates x^i with their associated Euclidean radius $r = \left(\sum_{i=1}^{3}(x^i)^2\right)^{1/2}$ and let δ_{ij} be the Euclidean metric components with respect to x^i. A three-dimensional manifold S is called *Euclidean at infinity*, if there exists a compact subset $\mathcal{K}$ of S such that $S \setminus \mathcal{K}$ is the disjoint union of a finite number of ends U_k. The initial data set $(S, h_{ij}, K_{ij}, \mu, j^i)$ is called *asymptotically flat* if S is Euclidean at infinity and at every end the metric h_{ij} and the tensor K_{ij} satisfy the following fall-off conditions

$$h_{ij} = \delta_{ij} + \gamma_{ij}\,, \quad K_{ij} = O(r^{-2})\,, \tag{18.4}$$

where $\gamma_{ij} = O(r^{-1})$, $\partial_k\gamma_{ij} = O(r^{-2})$, $\partial_l\partial_k\gamma_{ij} = O(r^{-3})$ and $\partial_k K_{ij} = O(r^{-3})$. These conditions are written in terms of Cartesian coordinates x^i attached at every end U_k. Here ∂_i denotes partial derivatives with respect to these coordinates.

At first sight it could appear that the notion of asymptotically flat manifold with *multiple ends* U_k is a bit artificial. Certainly, the most important case is when $S = \mathbb{R}^3$, for which this definition trivializes with $\mathcal{K} = B_R$ and only one end $U = \mathbb{R}^3 \setminus B_R$. Initial data for standard configurations of matter like stars or galaxies are modeled with $S = \mathbb{R}^3$. Also, gravitational collapse can be described with this kind of data. However, initial conditions with multiple ends and nontrivial interior $\mathcal{K}$ appear naturally in black hole initial data as we will see. In particular, the initial data for the Schwarzschild black hole has two asymptotic ends. On the other hand, this generalization does not imply any essential difficulty in the proofs of the theorems.

Only conditions on h_{ij} and K_{ij} are imposed in (18.4) and not on the matter fields μ and j^i, however since they are coupled by the constraint equations (18.1)–(18.2) the fall-off conditions (18.4) impose fall-off conditions also on (μ, j^i).

The fall-off conditions (18.4) are far from being the minimal requirements for the validity of the theorem. This is a rather delicate issue that has important consequences in the definition of the energy. We will discuss this point in Sect. 18.2. We have chosen these particular fall-of conditions because they are simple to present and they encompass a rich family of physical models.

For asymptotically flat initial data the total energy and linear momentum of the spacetime are defined as integrals over two-spheres at infinity at every end by the following expressions

$$E = \frac{1}{16\pi}\lim_{r\to\infty}\oint_{S_r}(\partial_j h_{ij} - \partial_i h_{jj})s^i\,ds_0\,, \tag{18.5}$$

$$P_i = \frac{1}{8\pi}\lim_{r\to\infty}\oint_{S_r}(K_{ik} - Kh_{ik})\,s^k\,ds_0\,, \tag{18.6}$$

where s^i is its exterior unit normal and ds_0 is the surface element of the two-sphere with respect to the Euclidean metric. We emphasize that for every end U_k we have a corresponding energy and linear momentum $E_{(k)}, P^i_{(k)}$, which can have different values. We will discuss examples of this in Sect. 18.2.

The quantities E and P_i are defined on the asymptotic ends and they depend only on the asymptotic behavior of the fields h_{ij} and K_{ij}. However, since h_{ij} and K_{ij} satisfy the constraint equations (18.1)–(18.2), and the dominant energy condition (18.3) holds, these quantities, in fact, carry information of the whole initial conditions.

The energy E and the linear momentum P_i are components of a four-vector $P_a = (E, P_i)$ (indices $a, b, c, \ldots$ are four-dimensional). We will discuss this in Sect. 18.3. The total mass of the spacetime is defined by

$$M = \sqrt{E^2 - P_iP_j\delta^{ij}}\,. \tag{18.7}$$

We have all the ingredients to present the positive energy theorem.

Theorem 18.1 Positive energy theorem
Let $(S, h_{ij}, K_{ij}, \mu, j^i)$ be an asymptotically flat (with many possible asymptotic ends), complete initial data set, such that the dominant energy condition (18.3) holds. Then the energy and linear momentum (E, P_i) defined by (18.5)–(18.6) satisfy

$$E \geq \sqrt{P_i P_j \delta^{ij}} \geq 0\,, \tag{18.8}$$

at every end. Moreover, $E = 0$ at any end if and only if the initial data correspond to the Minkowski spacetime.

The word *complete* means that (S, h_{ij}) as a Riemannian manifold is complete. That is, no singularities are present on the initial conditions. However, the spacetime can be singular since singularities can developed from regular initial conditions, for example, in the gravitational collapse. We will discuss this in more detail in Sect. 18.2.

Note that Theorem 18.1 allows the vector P^a to be null and nontrivial. However, it was shown in [18.1] that if the energy momentum vector P^a is null, then it vanishes identically.

One remarkable aspect of this theorem is that it is nontrivial even in the case where $S = \mathbb{R}^3$ and no matter fields $\mu = j^i = 0$ are present. This corresponds to the positivity of the energy of the pure vacuum gravitational waves. We present explicit examples of this in Sect. 18.2.

For spacetimes with black holes there are space-like surfaces that touch the singularity. For that kind of initial condition Theorem 18.1 does not apply. Physically it is expected that it should be possible to prove a positivity energy theorem for black holes without assuming anything about what happens inside the black hole. That is, it should be possible to prove an extension of the positive energy theorem for initial conditions with inner boundaries if the boundary represents a black hole horizon. The following theorem deals precisely with that problem.

Theorem 18.2 Positive energy theorem with black hole inner boundaries
Let (S, h_{ij}, K_{ij}) be an asymptotically flat, complete initial data set, with $S = \mathbb{R}^3 \setminus B$, where B is a ball. Assume that the dominant energy condition (18.3) holds and that ∂B is a black hole boundary. Then the energy momentum E, P^i defined by (18.5)–(18.6) satisfies

$$E \geq \sqrt{P^i P_i} \geq 0\,. \tag{18.9}$$

Moreover, $E = 0$ if and only if the initial data correspond to the Minkowski spacetime.

We will explain what are black hole inner boundary conditions in Sect. 18.2.

The plan of this chapter is the following. In Sect. 18.2 we discuss the concept of the energy E and present examples that illustrate the hypothesis of the positive energy theorem. In Sect. 18.3 we analyze the linear momentum P_i and describe its transformation properties. In Sect. 18.4 we review the main steps of the proof of theorems 18.1 and 18.2. Finally, in Sect. 18.5 other recent related results are discussed and the relevant current open problems are presented.

18.2 Energy

A remarkable feature of the asymptotic conditions (18.5) is that they imply that the total energy can be expressed exclusively in terms of the Riemannian metric h_{ij} of the initial data (and the linear momentum in terms of h_{ij} and the second fundamental form K_{ij}). Hence the notion of energy can be discussed in a pure Riemannian setting, without mentioning the second fundamental form. Moreover, as we will see, there is a natural corollary of the positive energy theorem for Riemannian manifolds. This corollary is relevant for several reasons. First, it provides a simpler and more relevant setting to prove the positive energy theorem. Second, and what is more important, it has surprising applications in other areas of mathematics. Finally, we will deal first with the Riemannian metric and then, in the next section, with the second fundamental form to incorporate the linear momentum, to reveal the different mathematical structures behind the energy concept.

In the previous section we introduced the notion of an end U, the energy was defined in terms of Riemannian metrics on U. To emphasize this important point we isolate the notion of energy defined in the theorems in the following definition.

Definition 18.1 Energy
Let h_{ij} be a Riemannian metric on an end U given in the coordinate system x^i associated with U. The energy is defined by

$$E = \frac{1}{16\pi} \lim_{r\to\infty} \oint_{S_r} (\partial_j h_{ij} - \partial_i h_{jj}) s^i \, \mathrm{d}s_0 . \tag{18.10}$$

Note that in this definition there is no mention of the constraint equations (18.1)–(18.2). Also, the definition only involves an end U, there are no assumptions on the interior of the manifold.

In the literature it is customary to call E the total mass and denote it by m or M. In this article, in order to emphasize that E is, in fact, the zero component of a four-vector we prefer to call it energy and reserve the name mass to the quantity M defined by (18.7). When the linear momentum is zero, both quantities coincide.

The definition of the total energy has three main ingredients: the end U, the coordinate system x^i, and the Riemannian metric h_{ij}. The metric is always assumed to be smooth on U; we will deal with singular metrics but these singularities will be in the interior region of the manifold and not on U.

There exist two potential problems with the definition 18.1. The first one is that the integral (18.10) could be infinite. The second, and more subtle, problem is that the mass seems to depend on the particular coordinate system x^i. Both problems are related with fall-of conditions for the metric. In the previous section, we introduced an example of this kind of condition in equation (18.4). These conditions are probably sufficient to model most physically relevant initial data. However, it is interesting to study the optimal fall-of conditions that are necessary to have a well-defined notion of energy and such that the energy is independent of the coordinate system.

To study this problem, we first introduce a general class of fall-of conditions as follows. Given an end U with coordinates x^i and an arbitrary real number α, we say that the metric h_{ij} on U is *asymptotically flat of degree* α if the components of the metric with respect to these coordinates have the following fall off in U as $r \to \infty$

$$h_{ij} = \delta_{ij} + \gamma_{ij} , \tag{18.11}$$

with $\gamma_{ij} = O(r^{-\alpha})$ and $\partial_k \gamma_{ij} = O(r^{-\alpha-1})$. The subtle point is to determine the appropriate α decay. To understand the meaning of this coefficient let us discuss the following relevant example given in [18.2] (see also [18.3]). Take the Euclidean metric δ_{ij} in Cartesian coordinates x^i and consider coordinates y^i defined by

$$y^i = \frac{\rho}{r} x^i , \tag{18.12}$$

where ρ is defined by

$$r = \rho + c\rho^{1-\alpha} , \tag{18.13}$$

for some constants c and α.

Note that $\rho = \left(\sum_{i=1}^{3} (y^i)^2\right)^{1/2}$. The components g'_{ij} of the Euclidean metric in coordinates y^i have the following form

$$g'_{ij} = \delta_{ij} + \gamma_{ij} , \tag{18.14}$$

where γ_{ij} satisfies the decay conditions (18.11) with the arbitrary α prescribed in the coordinate definition (18.13). That is, the metric in the new coordinate system y^i is asymptotically flat of degree α.

We calculate the energy in the coordinates y^i using the definition (18.10). We obtain

$$E = \begin{cases} \infty , & \alpha < 1/2 , \\ c^2/8 , & \alpha = 1/2 , \\ 0 , & \alpha > 1/2 . \end{cases} \tag{18.15}$$

Of course, we expect that the energy of the Euclidean metric should be zero in any coordinate system. The interesting point of this example is the limit case $\alpha = 1/2$; the example shows that if the energy has any chance to be coordinate independent, then we should impose $\alpha > 1/2$. The following theorem, proved in [18.4, 5], says that this condition is also sufficient.

Theorem 18.3
Let U be an end with a Riemannian metric h_{ij} such that is satisfies the fall-of conditions (18.11) with $\alpha > 1/2$. Also assume also that the scalar curvature R is integrable in U, that is,

$$\int_U |R| \, \mathrm{d}v < \infty . \tag{18.16}$$

Then the energy defined by (18.10) is unique and it is finite.

In this theorem unique means that if we calculate the energy in any coordinate system for which the metric satisfies the decay conditions (18.11) with $\alpha > 1/2$, we obtain the same result. This theorem ensures that the energy is a geometrical invariant of the Riemannian metric in the end U. Historically, this theorem was proved after the positive energy theorems. In the original proofs of the positive energy theorems different decay conditions for the metric were used. The decay conditions are usually formulated in terms of integrals of derivatives (i. e., Sobolev spaces) [18.4], which are more flexible for many applications. This particular formulation (which is simpler to present) of theorem 18.3 was taken from [18.6]. The decay conditions with $\alpha > 1/2$, together with the condition (18.16) on the scalar curvature, are called *mass decay conditions.* The freedom in the coordinates x^i is only a rigid motion at infinity [18.4].

Theorem 18.3 completes the geometric characterization of the energy at the end U. We now turn to positivity. It is clear that the energy can have any sign on U. The model example is given by the initial data for the Schwarzschild black hole, with metric on U given by

$$h_{ij} = \psi^4 \delta_{ij} \,, \tag{18.17}$$

where ψ is the following function

$$\psi = 1 + \frac{C}{2r} \,, \tag{18.18}$$

with C an arbitrary constant. Computing the energy for this metric we obtain $E = C$. The constant C can, of course, have any sign. It is, however, important to emphasize that theorem 18.3 asserts that the energy is well defined and it is an invariant of the geometry of the end, even when it is negative.

To ensure the positivity of the energy we need to impose two important conditions. One is a local condition: the positivity of the local energy given by the dominant energy condition (18.3). The other is a global condition on the manifold: the manifold should be complete or should have black hole boundaries.

Initial conditions with

$$K_{ij} = 0 \,, \tag{18.19}$$

are called time symmetric initial data. That is, time symmetric initial data are characterized only by a Riemannian metric h_{ij}. Conversely, any Riemannian metric can be interpreted as time symmetric initial data. However, an arbitrary metric will not satisfy the dominant energy condition (18.3). In effect, inserting condition (18.19) in the constraint equation (18.2) and using the dominant energy condition (18.18), we obtain

$$R \geq 0 \,. \tag{18.20}$$

Only metrics that satisfy (18.20) can be interpreted as time symmetric initial data for which the dominant energy condition holds. But then, any metric such that (18.20) holds satisfies the dominant energy condition and is a good candidate for the positive energy theorem. Hence we obtain the following corollary of theorem 18.1.

Corollary 18.1 Riemannian positive mass theorem
Let (S, h_{ij}) be a complete, asymptotically flat, Riemannian manifold. Assume that the scalar curvature is non-negative (i. e., condition (18.20)). Then the energy is non-negative at every end and it is zero at one end if and only if the metric is flat.

This corollary was proved with the optimal decay conditions for the metric in [18.4, 7].

The interesting mathematical aspect of this corollary is that there is no mention of the constraint equations, the second fundamental form, or the matter fields. This theorem is a result in pure Riemannian geometry. It has surprising applications in the solution of the Yamabe problem (see the review article [18.7] and references therein).

Note that it is not necessary to impose that the whole second fundamental form is zero to have (18.20); from equation (18.2) it is clear that it is enough to have $K = 0$. This class of initial data are called maximal and they have important properties (Chap. 16). In particular, positive energy theorems for this kind of data are easier to prove (mainly because condition (18.20) holds) than for general initial data.

Let us discuss some examples of corollary 18.1. We begin with the case with one asymptotic end and trivial topology, namely $S = \mathbb{R}^3$. For arbitrary functions ψ, metrics of the form (18.17) are called conformally flat; they provide a very rich family of initial conditions which have many interesting applications (for example, initial data for black hole collisions, see the review article [18.8]). The scalar curvature for this class of metrics is given by

$$R = -8\psi^{-5}\Delta\psi \,, \tag{18.21}$$

where Δ is the Euclidean Laplacian. If ψ satisfies the fall-of conditions

$$\psi = 1 + u\,, \quad u = O(r^{-1})\,, \quad \partial_k u = O(r^{-2})\,, \tag{18.22}$$

then the energy for this class of metric is given by

$$E = -\frac{1}{2\pi} \lim_{r\to\infty} \oint_{S_r} \partial_r \psi \, \mathrm{d}s_0\,. \tag{18.23}$$

For ψ given by (18.18) we obtain $R = 0$, and then the metric satisfies the local condition (18.20) for any choice of the constant C. However, this metric cannot be extended to $\mathbb{R}^3$ since the function ψ is singular at $r = 0$ and hence, as expected, Corollary 18.1 does not apply to this case. Let us try to prescribe a function with the same decay (and hence identical energy) but such that it is regular at $r = 0$. For example,

$$\psi = 1 + \frac{C}{2\sqrt{r^2 + C^2}}\,. \tag{18.24}$$

Using (18.23) we again obtain that $E = C$. For any value of C the function ψ is strictly positive and bounded on $\mathbb{R}^3$, and hence the metric is smooth on $\mathbb{R}^3$. That is, it satisfies the completeness assumption in corollary 18.1. Using (18.21) we compute the scalar curvature

$$R = 12\psi^{-5} \frac{C^3}{(r^2 + C^2)^{5/2}}\,. \tag{18.25}$$

We have $R \geq 0$ if and only if $C \geq 0$. Also, in this example the mass is zero if and only if the metric is flat.

Other interesting examples can be constructed with conformally flat metrics as follows. Let ψ be a solution of the Poisson equation

$$\Delta\psi = -2\pi\tilde{\mu}\,, \tag{18.26}$$

which satisfies the decay conditions (18.22), where $\tilde{\mu}$ is a non-negative function of compact support in $\mathbb{R}^3$. Solution of (18.26) can be easily constructed using the Green function of the Laplacian. By equation (18.21), the scalar curvature of the associated conformal metric (18.17) will be non-negative and the function $\tilde{\mu}$ is related to the matter density μ by

$$\mu = \frac{R}{16\pi} = \tilde{\mu}\psi^{-5}\,. \tag{18.27}$$

Note that, in this example, we cannot prescribe exactly the matter density μ; we prescribe a conformal rescaling of μ. However, it is enough to control the support of μ. The support of μ represents the localization of the matter sources. Outside the matter sources the scalar curvature (for time symmetric data) is zero.

For conformally flat metrics in $\mathbb{R}^3$ there is a very simple proof of corollary 18.1. We write equation (18.21) as

$$\frac{R}{8} = -\partial^i\left(\frac{\partial_i\psi}{\psi^5}\right) - 5\frac{|\partial\psi|^2}{\psi^6}\,. \tag{18.28}$$

Integrating this equation in $\mathbb{R}^3$, using the Gauss theorem for the first term on the right-hand side, the condition $\psi \to 1$ as $r \to \infty$ and the expression (18.23) for the energy, we finally obtain

$$E = \frac{1}{2\pi} \int_{\mathbb{R}^3} \left(\frac{R}{8} + 5\frac{|\partial\psi|^2}{\psi^6}\right) \mathrm{d}v_0\,, \tag{18.29}$$

where $\mathrm{d}v_0$ is the flat volume element. This formula proves that for metric of the form (18.17) we have $E \geq 0$ if $R \geq 0$ and $E = 0$ if and only if $h_{ij} = \delta_{ij}$. This proof easily generalizes to conformally flat maximal initial data.

Asymptotically flat initial conditions in $\mathbb{R}^3$ with no matter sources (i. e., $\mu = j^i = 0$) represent pure gravitational waves. They are conceptually important because they describe the dynamic of pure vacuum, independent of any matter model. Note that in that case the energy condition (18.3) is trivially satisfied.

In the previous examples the only solution with pure vacuum $R = 0$ in $\mathbb{R}^3$ is the flat metric, because by equation (18.21) we obtain $\Delta\psi = 0$, and the decay condition (18.22) implies $\psi = 1$. In order to construct pure waves initial data we allow for a more general kind of conformal metrics, let h_{ij} be given by

$$h = \mathrm{e}^{\sigma}\left[\mathrm{e}^{-2q}(\mathrm{d}\rho^2 + \mathrm{d}z^2) + \rho^2\,\mathrm{d}\varphi^2\right], \tag{18.30}$$

where (ρ, z, φ) are cylindrical coordinates in $\mathbb{R}^3$ and the functions q and σ depend only on (ρ, z). That is, the metric h_{ij} given by (18.30) is axially symmetric.

The scalar curvature of the metric (18.30) is given by

$$-\frac{1}{8}R\mathrm{e}^{(\sigma-2q)} = \frac{1}{4}\Delta\sigma + \frac{1}{16}|\partial\sigma|^2 - \frac{1}{4}\Delta_2 q\,, \tag{18.31}$$

where Δ, as before, is the three-dimensional flat Laplacian and Δ_2 is the two-dimensional Laplacian in cylin-

drical coordinates given by

$$\Delta_2 q = \partial^2_\rho q + \partial^2_z q\,. \tag{18.32}$$

If we impose $R = 0$, equation (18.31) reduces to

$$\Delta\psi - \frac{1}{4}\Delta_2 q = 0\,, \tag{18.33}$$

where $\psi^4 = e^\sigma$. To construct metrics of the form (18.30) that satisfies $R = 0$ a function q is prescribed and then the linear equation (18.33) is solved for ψ. The function q cannot be arbitrary, it should satisfy a global condition (which is related with the Yamabe problem mentioned above), see [18.9] for details. This kind of metric is called Brill waves. These were used by *D. Brill* in one of the first proofs of the positive energy theorem [18.10]. Let us discuss this proof.

In order to be smooth at the axis the metric (18.30) should satisfy $q = 0$ at $\rho = 0$. For simplicity we also impose a strong fall-of condition on q at infinity, namely $q = O(r^{-2})$, $\partial_i q = O(r^{-2})$. For σ we impose $\sigma = O(r^{-1})$ and $\partial_i\sigma = O(r^{-2})$. Using these decay assumptions it is straightforward to check that the energy of the metric (18.30) is given by

$$E = -\frac{1}{8\pi}\lim_{r\to\infty}\oint_{S_r}\partial_r\sigma\,\mathrm{d}s_0\,. \tag{18.34}$$

With the Gauss theorem, using that $q = 0$ at the axis and the fall-of condition of q at infinity, we obtain that

$$\int_{\mathbb{R}^3}\Delta_2 q\,\mathrm{d}v_0 = 0\,. \tag{18.35}$$

Integrating equation (18.31) in $\mathbb{R}^3$, using (18.35) and using the expression (18.34) for the energy we obtain

$$E = \frac{1}{8\pi}\int_{\mathbb{R}^3}\left(\frac{1}{2}|\partial\sigma|^2 + Re^{\sigma-2q}\right)\mathrm{d}v_0\,. \tag{18.36}$$

That is, $R \geq 0$ implies $E \geq 0$. In particular for vacuum $R = 0$, we have

$$E = \frac{1}{16\pi}\int_{\mathbb{R}^3}|\partial\sigma|^2\,\mathrm{d}v_0\,. \tag{18.37}$$

This positivity proof can be extended in many ways, in particular it has applications for the inequality between energy and angular momentum discussed in Sect. 18.5 (see the review article [18.11] and the lectures notes [18.6, 12], and references therein).

We turn now to manifolds with many asymptotic flat ends and interior $\mathcal{K}$ with non-trivial topology defined in Sect. 18.1. Let us first present some basic example of the definition of asymptotic Euclidean manifold, without mentioning the metric.

Taking out a point in $\mathbb{R}^3$, the manifold $S = \mathbb{R}^3 \setminus \{0\}$ is asymptotic Euclidean with two ends, which we denote by U_0 and U_1. In effect, let B_2 and B_1 be two balls centered at the origin with radius 2 and 1, respectively. Define $\mathcal{K}$ as the annulus centered at the origin $B_2 \setminus B_1$. Then $S \setminus \mathcal{K}$ has two components U_0 and U_1, where $U_0 = \mathbb{R}^3 \setminus B_2$ and $U_1 = B_1 \setminus \{0\}$. The set U_0 is clearly an end. The set U_1 is also an end since a ball minus a point is diffeomorphic to $\mathbb{R}^3$ minus a ball. This can be explicitly seen using Cartesian coordinates centered at the origin x^i; then the map given by the inversion

$$y^i = r^{-2}x^i\,, \tag{18.38}$$

provides the diffeomorphism between $\mathbb{R}^3 \setminus B_1$ and $B_1 \setminus \{0\}$.

In the same way $\mathbb{R}^3$ minus a finite number N of points i_k is a Euclidean manifold with $N + 1$ ends. For each i_k take a small ball B_k of radius $r_{(k)}$, centered at i_k, where $r_{(k)}$ is small enough such that B_k does not contain any other $i_{k'}$ with $k' \neq k$. Take B_R, with large R, such that B_R contains all points i_k. The compact set $\mathcal{K}$ is given by $\mathcal{K} = B_R \setminus \sum_{k=1}^N B_k$ and the open sets U_k are given by $B_k \setminus i_k$, for $1 \leq k \leq N$, and U_0 is given by $\mathbb{R}^3 \setminus B_R$.

Another example is a torus $\mathbb{T}^3$ minus a point i_0. Take a small ball B centered at i_0. Then the manifold is asymptotic Euclidean with $\mathcal{K} = \mathbb{T}^3 \setminus B$ and only one end $U = B \setminus i_0$. This is an example of a Euclidean manifold with one asymptotic end but non-trivial $\mathcal{K}$. More generally, given any compact manifold, if we subtract a finite number of points we obtain an asymptotically Euclidean manifold with multiple ends. Note that the topology of the compact core $\mathcal{K}$ can be very complicated.

Let us consider now Riemannian metrics on these asymptotic Euclidean manifolds. Consider the manifold $S = \mathbb{R}^3 \setminus \{0\}$ and the metric given by (18.17) and (18.18). The function ψ is smooth on S for any value of the constant C, however if $C < 0$ then ψ vanishes at $r = -2/C$, and hence the metric is not defined at those points. That is, the metric h_{ij} is smooth on S only when $C \geq 0$. We have seen that S has two asymptotic

ends, let us check that the metric h_{ij} is asymptotically flat (i. e., that it satisfies the decay conditions (18.4)) at both ends U_0 and U_1. On U_0, the metric in the coordinates x^i is clearly asymptotically flat. However, note that in these coordinates the metric is not asymptotically flat at the end U_1 (which, in these coordinates is represented by a neighborhood of $r = 0$), in fact, the components of the metric are singular at $r = 0$. However, using a coordinate an inversion of coordinates like (18.38) is straightforward to prove that the metric is asymptotically flat also at $r = 0$. More precisely, consider the coordinate transformation

$$y^i = \left(\frac{C}{2}\right)^2 \frac{1}{r^2} x^i , \quad \rho = \left(\frac{C}{2}\right)^2 \frac{1}{r} . \tag{18.39}$$

In terms of these coordinates the metric has the form

$$h'_{ij} = \left(1 + \frac{C}{2\rho}\right)^4 \delta_{ij} . \tag{18.40}$$

We have chosen the constant factor in the coordinate transformation (18.40) in such a way that the transformation it is, in fact, the well known isometry of this metric, this choice, however, is not essential. The metric (18.40) is clearly asymptotically flat at U_1. Note that we have two energies, one for each end; the two are equal and given by the constant C. In this example, the positivity of the mass is enforced purely by the global requirement of completeness of the metric (the energy condition is satisfied for arbitrary C). It is this condition that fails when $C < 0$. In that case the metric is defined on a manifold with boundary $S = \mathbb{R}^3 \setminus B_{-2/C}$, and the metric vanishes at the boundary $\partial B_{-2/C}$. In particular, the two-surface $\partial B_{-2/C}$ has zero area. This motivated the concept of *zero area singularities* introduced in [18.13], where interesting results are presented concerning negative energy defined on this class of singular metrics.

In the previous example the energies at the different ends are equal. It is straightforward to construct an example for which the two energies are different. Consider the following function

$$\psi = 1 + \frac{C}{2r} + g , \tag{18.41}$$

where g is a smooth function on $\mathbb{R}^3$ such that $g = O(r^{-2})$ as $r \to \infty$ and $g(0) = a$. Making the same calculation we obtain that the energy at one end is $E_0 = C$ (here we use the decay conditions on g, otherwise the function g will contribute to the energy at that end). However, at the other end the components of the metric in the coordinates y^i are given by

$$h'_{ij} = \left(1 + \frac{C(1+g)}{2\rho}\right)^4 \delta_{ij} , \tag{18.42}$$

and hence we have that

$$E_1 = C(1+a) . \tag{18.43}$$

Note that in order to satisfy the energy condition (18.20) g (and hence a) cannot be arbitrary, we must impose the following condition on g

$$\Delta g \le 0 . \tag{18.44}$$

Using (18.44), the decay assumption on g and the maximum principle for the Laplacian (see, for example, the version of the maximum principle in the Appendix of [18.14]) it is easy to prove that $g \ge 0$ and then $a \ge 0$.

Consider the manifold $S = \mathbb{R}^3 \setminus \{i_1\}, \{i_2\}$ with three asymptotic ends. Also consider the function given by (this nice example was constructed in [18.15])

$$\psi = 1 + \frac{C_1}{2r_1} + \frac{C_2}{2r_2} , \tag{18.45}$$

where r_1 and r_2 are the Euclidean radius centered at the points i_1 and i_2, respectively, and C_1 and C_2 are constant. Note that $\Delta\psi = 0$ and hence the metric defined by (18.17) has $R = 0$. As before, only when $C_1, C_2 \ge 0$ the metric is smooth on S. Also, using a similar calculation as in the case of two ends it is not difficult to check that the metric is asymptotically flat on the three ends. Moreover, the energies of the different ends are given by

$$E_0 = C_1 + C_2 , \quad E_1 = C_1 + \frac{C_1 C_2}{L} ,$$
$$E_2 = C_2 + \frac{C_1 C_2}{L} , \tag{18.46}$$

where L is the Euclidean distance between i_1 and i_2. We see that they are all positive and, in general, different. These initial conditions model a head on collision of two black holes and they have been extensively used in numerical simulations of black hole collisions (see, for example, [18.16] and references therein).

We analyze the case of the wormhole [18.17], which is an example of $\mathcal{K}$ with more complicated topology.

Consider the metric on the compact manifold $S = \mathbb{S}^1 \times \mathbb{S}^2$ given by

$$\gamma = \mathrm{d}\mu^2 + (\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2)\,, \tag{18.47}$$

where the coordinate ranges are $-\pi < \mu \leq \pi$, and the sections $\mu = const$ are two-spheres. Let h_{ij} be given by

$$h_{ij} = \psi^4 \gamma_{ij}\,, \tag{18.48}$$

where the function ψ is

$$\psi = \sum_{n=-\infty}^{n=\infty} [\cosh(\mu + 2n\pi)]^{-1/2}\,. \tag{18.49}$$

This function blows up at $\mu = 0$. Hence, the metric h_{ij} is defined on S minus the point $\mu = 0$. We have seen that this is an asymptotic Euclidean manifold with one asymptotic end. It can be proved that the metric is asymptotically flat at that end (see [18.17] for details). Also, the function ψ is chosen in such a way that the scalar of the h_{ij} curvature vanished. Moreover, the energy is given by

$$E = 4\sum_{n=1}^{\infty}(\sinh(n\pi))^{-1}\,, \tag{18.50}$$

which is positive.

Finally, consider the initial data for the Reissner–Nördstrom black hole given by a metric of the form (18.17) with ψ given by

$$\psi = \frac{1}{2r}\sqrt{(q + 2r + C)(-q + 2r + C)}\,, \tag{18.51}$$

where C and q are constant. The scalar curvature of this metric is given by

$$R = \frac{2q^2}{\psi^8 r^4}\,. \tag{18.52}$$

Which is non-negative for any value of the constants. When $C > |q|$, the metric is asymptotically flat with two ends U_0 and U_1 as in the example (18.18). The energy on both ends is given by $E = C$. The positive energy theorem applies to this case. If $C < |q|$ then the metric is singular, there is only one end U_0, and the energy on that end is given by C. Note that in this case it is still possible to have positive energy $0 < C < |q|$, but the positive energy theorem does not apply because it is a singular metric. The borderline case $C = |q|$ represents the extreme black hole. The manifold is $\mathbb{R}^3$ minus a point and the metric is smooth on that manifold. However, the metric is asymptotically flat only at the end U_0, on the other end it is asymptotically cylindrical. Hence this version of the positive energy theorem does not apply for these data. The asymptotically cylindrical end is a feature of all extreme black holes. For discussions on this kind of geometry, see [18.11] and references therein.

So far, we have discussed complete manifolds without boundaries or manifolds with boundaries in which the metric is singular at the boundaries. We now analyze the important case of black hole boundaries.

Black hole boundaries are defined in terms of marginally trapped surfaces. A marginally trapped surface is a closed two-surface such that the outgoing null expansion Θ_+ vanishes (more details on this important concept can be seen in [18.18]). If such a surface is embedded on a space-like three-dimensional surface, then the expansion Θ_+ can be written in terms of the initial conditions as follows

$$\Theta_+ = H - K_{ij}s^i s^j + K\,, \tag{18.53}$$

where

$$H = D_i s^i\,, \tag{18.54}$$

is the mean curvature of the surface. Here s^i is the unit normal vector to the surface. For time symmetric initial data, condition $\Theta_+ = 0$ reduces to

$$H = 0\,. \tag{18.55}$$

Surfaces that satisfy condition (18.55) are called minimal surfaces, because (18.55) is satisfied if and only if the first variation of the area of the surface vanishes. These kinds of surfaces have been extensively studied in Riemannian geometry (see the book [18.19] for an introduction to the subject). We have seen that a marginally trapped surface on time symmetric initial data is a minimal surface. That is, black hole boundaries translate, for these kinds of data, into a pure Riemannian boundary condition. Then, we have the following corollary of theorem 18.2.

Corollary 18.2 Black holes in Riemannian geometry
Let (S, h_{ij}) be a complete, asymptotically flat, Riemannian manifold with compact boundary. Assume that the scalar curvature is non-negative (i. e., condition (18.20)) and that the boundary is a minimal surface (i. e., it satisfies (18.55)). Then the energy is non-negative and it is zero at one end if and only if the metric is flat.

Let us give a very simple example that illustrate this theorem. Consider the function ψ given by (18.18). It is well known that the surface $r = C/2$ is a minimal surface (it represents the intersection of the Schwarzschild black hole event horizon with the spacetime surface $t = constant$ in Schwarzschild coordinates). To verify, that we compute H for the two-surfaces $r = constant$ for the metric (18.17). The unit normal vector is given by

$$s^i = \psi^{-2}\left(\frac{\partial}{\partial r}\right)^i . \tag{18.56}$$

Then we have

$$H = D_i s^i = \frac{4}{\psi^3}\left(\partial_r \psi + \frac{\psi}{2r}\right). \tag{18.57}$$

Then condition (18.55) is equivalent to

$$0 = \partial_r \psi + \frac{\psi}{2r} = \frac{1}{2r} - \frac{C}{4r^2} , \tag{18.58}$$

and hence for $r = C/2$ we have a minimal surface. Note that C must be positive in order to have a minimal surface. Previously we discussed this example in the complete manifold, without boundaries, $\mathbb{R}^3 \setminus \{0\}$. In that case corollary 18.1 applies. We can also consider the same metric but in the manifold with boundary $Rt \setminus B_{C/2}$. Since we have seen that $\partial B_{C/2}$ is a minimal surface, then corollary 18.2 applies to that case. To emphasize the scope of this corollary, we slightly extend this example in the following form. Consider ψ given by

$$\psi = \left(1 + \frac{C}{2r}\right)\chi(r) , \tag{18.59}$$

where $\chi(r)$ is a function such that is $\chi = 1$ for $r > C/2$ and arbitrary for $r < C/2$. Corollary 18.2 applies to this case, since again the boundary is a minimal surface. Note that inside the minimal surface the function χ is arbitrary, in particular it can blow up and it does not need to satisfy the energy condition. Corollary 18.1 certainly does not apply to this case.

18.3 Linear Momentum

The total mass M defined by (18.7) in terms of the energy and linear momentum (18.5)–(18.6) represents the total amount of energy of the spacetime. The first basic question we need to address is in what sense M is independent of the choice of initial conditions that describe the same spacetime. That is, given a fixed spacetime we can take different space-like surfaces on it, on each surface we can calculate the initial data set and hence we have a corresponding M, do we obtain the same result? We will see that the answer to this question strongly depends on the fall-of conditions (18.4).

To illustrate that, let us consider the Schwarzschild spacetime. We recall that in the following examples the spacetime is fixed and we only chose different space-like surfaces on it. The spacetime metric is given in Schwarzschild coordinates (t, r_s, θ, ϕ) by

$$\mathrm{d}s^2 = -\left(1 - \frac{2C}{r_s}\right)\mathrm{d}t^2 + \left(1 - \frac{2C}{r_s}\right)^{-1}\mathrm{d}r_s^2 + r_s^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2) . \tag{18.60}$$

These coordinates are singular at $r_s = 2C$ and hence they do not reveal the global structure of the surfaces $t =$ constant. The most direct way to see that these surfaces are complete three-dimensional manifolds is to use the isotropic radius r defined by

$$r_s = r\left(1 + \frac{C}{2r}\right)^2 . \tag{18.61}$$

In isotropic coordinates the line element is given by

$$\mathrm{d}s^2 = -\left(\frac{1 - \frac{C}{2r}}{1 + \frac{C}{2r}}\right)^2 \mathrm{d}t^2 + \left(1 + \frac{C}{2r}\right)^4 (\mathrm{d}r^2 + \mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2) . \tag{18.62}$$

The initial data on the slice $t =$ constant are given by

$$h_{ij} = \left(1 + \frac{C}{2r}\right)^4 \delta_{ij} , \quad K_{ij} = 0 . \tag{18.63}$$

These are the time symmetric initial data studied in Sect. 18.2. The linear momentum of these data is obviously zero, then the total mass M is equal to the

energy E calculated in the previous section, and we obtain the expected result $M = C$.

We take another foliation of space-like surfaces. We write the metric (18.62) in the Gullstrand–Painlevé coordinates $(t_{\text{gp}}, r_{\text{s}}, \theta, \phi)$ (see [18.20] and references therein). We obtain

$$\mathrm{d}s^2 = -\left(1 - \frac{2C}{r_{\text{s}}}\right)\mathrm{d}t_{\text{gp}}^2 + 2\sqrt{\frac{2C}{r_{\text{s}}}}\,\mathrm{d}t_{\text{gp}}\,\mathrm{d}r_{\text{s}} + \mathrm{d}r_{\text{s}}^2 + r_{\text{s}}^2\,\mathrm{d}\theta^2 + r_{\text{s}}^2 \sin^2\theta\,\mathrm{d}\phi^2 . \quad (18.64)$$

The slices $t_{\text{gp}} =$ constant in these coordinates have the following initial data

$$h_{ij} = \delta_{ij}\,, \quad K_{ij} = \frac{\sqrt{2m}}{r_{\text{s}}^{3/2}}\left(\delta_{ij} - \frac{3}{2}s_i s_j\right), \quad (18.65)$$

where s^i is the radial unit normal vector with respect to the flat metric δ_{ij}. We see that the intrinsic metric is flat and hence the energy E is clearly zero. The linear momentum is also zero, because if we calculate the integral (18.6) at a sphere of finite radius (note that the limit is in danger to diverge because the radial dependence of K_{ij} in (18.65)) the angular variables integrate to zero. Hence we obtain that for these surfaces the total mass M is zero. What happens is that the second fundamental form (18.65) does not satisfy the decay condition (18.4) since it falls off like $O(r^{-3/2})$. It can be proved that any initial conditions that satisfy (18.4) in the same spacetime give the same total mass M.

We consider another foliation which reveals the Lorentz transformation properties of (E, P^i). Let (x, y, z) be the associated Cartesian coordinates of the isotropic coordinates (r, θ, ϕ), that is,

$$x = r\cos\phi\sin\theta\,,$$
$$y = r\sin\phi\cos\theta\,,$$
$$z = r\cos\theta\,. \quad (18.66)$$

We consider the line element (18.62) written in terms of the coordinates (t, x, y, z) and we perform the following change of coordinates which represents a boost in the z direction

$$\hat{t} = \gamma^{-1}(t - vz)\,, \quad (18.67)$$
$$\hat{z} = \gamma^{-1}(-vt + \hat{z})\,, \quad (18.68)$$
$$\hat{x} = x\,, \quad (18.69)$$
$$\hat{y} = y\,, \quad (18.70)$$

where v is a constant and $\gamma = \sqrt{1 - v^2}$. Consider a surface $\hat{t} =$ constant in these coordinates. The intrinsic metric is given by

$$h = \psi^4(\mathrm{d}\hat{x}^2 + \mathrm{d}\hat{y}^2) + \gamma^{-2}(-N^2v^2 + \psi^4)\,\mathrm{d}\hat{z}^2\,, \quad (18.71)$$

where

$$\psi = 1 + \frac{C}{2r}\,, \quad N = \frac{1 - \frac{C}{2r}}{1 + \frac{C}{2r}}\,. \quad (18.72)$$

The radius r can be written in terms of the hat coordinates as follows

$$r = \sqrt{x^2 + y^2 + z^2} = \sqrt{\hat{x}^2 + \hat{y}^2 + \gamma^{-2}(v\hat{t} + \hat{z})^2}\,. \quad (18.73)$$

One can check that the metric h given by (18.71) is asymptotically flat in the coordinates $(\hat{x}, \hat{y}, \hat{z})$. Then, we can compute the energy of this metric and we obtain

$$E = \gamma^{-1}C\,. \quad (18.74)$$

To obtain the linear momentum we need to compute the second fundamental form of the slice. The calculations are long (see, for example, [18.21] for details), the final result is the following

$$P_x = 0\,, \quad P_y = 0\,, \quad P_z = vC\gamma^{-1}\,. \quad (18.75)$$

Using (18.74) and (18.75) we obtain

$$M = \sqrt{E^2 - P^iP^j\delta_{ij}} = C\,. \quad (18.76)$$

That is, the quantities E, P^i transform like a four-vector under asymptotic Lorentz transformations of coordinates.

18.4 Proof

In Sect. 18.2 we have presented two proofs of the positive energy theorem for two particular cases, for other proofs that apply to other relevant particular cases (like spherical symmetry and the weak field limit), see [18.6, 12, 22] and references therein.

The first general proof of the positive energy theorem was done by *Schoen* and *Yau* [18.23]. Shortly afterwards it was followed by a proof by Witten [18.24], who used completely different methods. The proof of the Penrose inequality by *Huisken* and *Illmanen* in [18.25] (we briefly discuss this work in Sect. 18.5) also provided a new proof of the positive energy theorem (which is based on an idea of *Geroch* [18.26])

The simplest of all these proofs is, by far, that of Witten. Moreover, it resembles other positivity proofs in physics: the total energy is written as a positive definite integral in the space. In this section we review this proof. The aim is to present all the relevant steps in the most elementary way.

This proof uses, in an essential way, spinors. We refer the reader to Chap. 15 for an introduction to this subject. We will follow the notation of that chapter in this section.

There exists various reformulations of the original proof by Witten; in this section we essentially follow references [18.27–30].

The proof uses only spinors defined on the space-like surface, however it is more transparent to begin with spinor fields in the spacetime and then, at the very end, to restrict them to the space-like surface. Also, this way of constructing the proof easily generalizes to the proof of the positivity of the energy at null infinity (Bondi mass) [18.30].

Let (M, g_{ab}) be a four-dimensional Lorentzian manifold with connection ∇_a. In this section we use the signature $(+---)$ to be consistent with the literature on spinors. Unfortunately this signature gives a negative sign to the Riemannian metrics on the space-like surfaces used in the previous sections.

Let λ_A be a spinor field in the spacetime; the spin connection is denoted by $\nabla_{AA'}$, and we use the standard notation $a = AA'$ to identify spinor indices with tensorial indices.

The proof of the positive energy theorem is based on the remarkable properties of a two-form $\mathbf{\Omega}$ called the Nester–Witten form [18.24, 31], defined as follows. The computations of this section involve integration on different kind of surfaces and hence it is convenient to use differential forms instead of ordinary tensors. We will denote them with boldface and no indices (for an introduction to forms see, for example, Appendix B in [18.18], we will follow the notation and convention of this reference).

Consider the following complex tensor

$$\Omega_{ab} = -\mathrm{i}\bar{\lambda}_{B'}\nabla_{AA'}\lambda_B\,. \qquad (18.77)$$

From this tensor we construct the complex 2-form $\mathbf{\Omega}$ by

$$\mathbf{\Omega} = \Omega_{[ab]}\,. \qquad (18.78)$$

Explicitly we have

$$\mathbf{\Omega} = \frac{\mathrm{i}}{2}\left(\bar{\lambda}_{A'}\nabla_{BB'}\lambda_A - \bar{\lambda}_{B'}\nabla_{AA'}\lambda_B\right). \qquad (18.79)$$

The forms used in the following are always tensor fields (usually complex) but they are constructed out of spinors, as in the case of $\mathbf{\Omega}$. In order to define these forms we need to antisymmetrize tensorial indices, to avoid a complicated notation we will always define first the tensor field in terms of spinors (as in equation (18.77)) and then define the form antisymmetrizing the tensor indices (as in (18.78)). When there are more than two tensorial indices the explicit expression of the differential form (like (18.79)) can be lengthy and it is not usually needed. The spinor λ_A has an associated (future directed) null vector ξ_a given by

$$\xi^a = \lambda^A\bar{\lambda}^{A'}\,. \qquad (18.80)$$

Note that the $\mathbf{\Omega}$ cannot be written in terms of derivatives of pure tensors fields like ξ_a and ∇_a.

The strategy of the proof is the following. Consider the exterior derivative $\mathrm{d}\mathbf{\Omega}$ (which is a three-form) and integrate it on a space-like, asymptotically flat, three-surface S. Using Stoke's theorem we obtain

$$\sum_k \lim_{r\to\infty} \oint_{S_r} \mathbf{\Omega} = \int_S \mathrm{d}\mathbf{\Omega}\,. \qquad (18.81)$$

We assume that S is an asymptotically Euclidean manifold with k asymptotic ends U_k. The two-form $\mathbf{\Omega}$ has two important properties. The first one is that the left-hand side of (18.81) gives the total energy-momentum of a prescribed asymptotic end. The second is that the integrand of the right-hand side is non-negative. Both properties depend on the way in which the spinor field λ^A is prescribed.

We begin with the first property. Note that the integrand on the left-hand side of (18.81) is complex. However, the imaginary part of $\mathbf{\Omega}$ is given by

$$\mathbf{\Omega} - \bar{\mathbf{\Omega}} = \mathrm{i}\nabla_{[a}\xi_{b]} = i\mathrm{d}\boldsymbol{\xi} \ , \tag{18.82}$$

where, to be consistent with our notation, we write $\boldsymbol{\xi}$ for the the one-form ξ_a. That is, the imaginary part is the exterior derivative of a one-form and hence its integral over a closed two-surface is zero. Hence the boundary integral is always real, for arbitrary spinors λ^A.

To prove the desired property, we need to impose fall-of conditions on the spinor λ^A. Fix one arbitrary end k (from now on we will always work on that end and hence we suppress the label k). Let $\mathring{\lambda}^A$ be an arbitrary *constant* spinor; we require that the spinor λ^A satisfies on that end

$$\lambda^A = \mathring{\lambda}^A + \gamma^A \ , \quad \gamma^A = O(r^{-1}) \ . \tag{18.83}$$

We also assume that the partial derivatives of γ^A are $O(r^{-2})$ and we require that λ^A *decays to zero at every other end.*

The idea is to prove that at the chosen end we have

$$P_a\mathring{\xi}^a = \frac{1}{8\pi}\lim_{r\to\infty}\oint_{S_r}\mathbf{\Omega} \ , \tag{18.84}$$

where $P_a = (E, P_i)$, with E and P_i defined by (18.5)–(18.6), and $\mathring{\xi}^a$ is the constant null vector determined by the constant spinor $\mathring{\lambda}^A$ by

$$\mathring{\xi}^a = \mathring{\lambda}^A\mathring{\bar{\lambda}}^{A'} \ . \tag{18.85}$$

Note that the boundary integral on the right-hand side of (18.84) determines both the energy and the linear momentum of the end.

To prove (18.84) the most important step is to prove that the value of the integral depends only on the constant spinor $\mathring{\lambda}^A$ and not on γ^A. We emphasize, as we will see, that a naive counting of the fall behavior of the different terms in $\mathbf{\Omega}$, under the assumption (18.83), does not prove this result. Using the decomposition (18.83) we write $\mathbf{\Omega}$ as

$$\mathbf{\Omega} = \mathring{\mathbf{\Omega}} + \mathbf{\Gamma} \ , \tag{18.86}$$

where

$$\mathring{\Omega}_{ab} = -\mathrm{i}\mathring{\bar{\lambda}}_{B'}\nabla_{AA'}\mathring{\lambda}_B \ , \quad \mathring{\mathbf{\Omega}} = \mathring{\Omega}_{[ab]} \ , \tag{18.87}$$

and

$$\Gamma_{ab} = -\mathrm{i}\left(\mathring{\bar{\lambda}}_{B'}\nabla_{AA'}\gamma_B + \bar{\gamma}_{B'}\nabla_{AA'}\mathring{\lambda}_B + \bar{\gamma}_{B'}\nabla_{AA'}\gamma_B\right) ,$$
$$\mathbf{\Gamma} = \Gamma_{[ab]} \ . \tag{18.88}$$

That is, $\mathring{\mathbf{\Omega}}$ depends only on $\mathring{\lambda}^A$.

We would like to prove that $\mathbf{\Gamma} = O(r^{-3})$ and hence it does not contribute to the integral at infinity (18.84). Consider the third term in (18.88). The covariant derivative $\nabla_{AA'}\gamma_B$ has two terms, the first one contains partial derivatives of γ_B which, by assumption, are $O(r^{-2})$. The second term contains products of γ_B and the connections coefficients of the spacetime metric g_{ab} evaluated at the asymptotic end of the space-like surface S. These coefficients are first derivatives of g_{ab}, they can be written as first derivatives of the intrinsic Riemannian metric and the second fundamental form of the surfaces and hence, by assumption (recall that S is asymptotically flat and hence we have the fall-of conditions (18.4)) they are $O(r^{-2})$. We conclude that $\nabla_{AA'}\gamma_B = O(r^{-2})$ and hence $\bar{\gamma}_{B'}\nabla_{AA'}\gamma_B = O(r^{-3})$. We proceed in a similar way for the second term: since $\mathring{\lambda}^A$ is constant the covariant derivative $\nabla_{AA'}\mathring{\lambda}_B$ contains connection coefficients times constants and hence we have $\nabla_{AA'}\mathring{\lambda}_B = O(r^{-2})$, and then $\bar{\gamma}_{B'}\nabla_{AA'}\mathring{\lambda}_B = O(r^{-3})$. However, using the same argument we obtain that the first term in (18.88) is $O(r^{-2})$ and then it can contribute to the integral. However, we can re-write Γ_{ab} as follows

$$\Gamma_{ab} = -\mathrm{i}\left(\nabla_{BB'}\left(\gamma_A\mathring{\bar{\lambda}}_{A'}\right) - \gamma_A\nabla_{BB'}\mathring{\bar{\lambda}}_A + \bar{\gamma}_{B'}\nabla_{AA'}\mathring{\lambda}_B + \bar{\gamma}_{B'}\nabla_{AA'}\gamma_B\right) . \tag{18.89}$$

The first term in (18.89), which is the problematic one, contributes to $\mathbf{\Gamma}$ with the derivative of a one-form, and hence it integrates to zero over a closed two-surface. The new second term in (18.89) is clearly $O(r^{-3})$. We have proved that

$$\lim_{r\to\infty}\oint_{S_r}\mathbf{\Omega} = \lim_{r\to\infty}\oint_{S_r}\mathring{\mathbf{\Omega}} \ . \tag{18.90}$$

Note that $\mathring{\mathbf{\Omega}}$ is $O(r^{-2})$, and hence the integral converges. Also, the asymptotic value of $\mathring{\mathbf{\Omega}}$ at infinity contains a combination of the first derivative of the intrinsic metric and the second fundamental form of the sur-

face S multiplied by the constants $\mathring{\lambda}^A$. It can be proved, essentially by an explicit calculation, that this combination is precisely $P_a\xi^a$ (see [18.24, 31] and also [18.32]).

We turn to the second property of $\mathbf{\Omega}$. Recall that the exterior derivative of a p-form is given by

$$\mathrm{d}\mathbf{\Omega} = (p+1)\nabla_{[a}\Omega_{b_1\cdots b_p]} \,. \tag{18.91}$$

We have

$$\mathrm{d}\mathbf{\Omega} = \boldsymbol{\alpha} + \boldsymbol{\beta} \,, \tag{18.92}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the following three-forms

$$\alpha_{abc} = -\mathrm{i}\bar{\lambda}_{C'}\nabla_a\nabla_b\lambda_C \,, \quad \boldsymbol{\alpha} = \alpha_{[abc]} \,, \tag{18.93}$$

and

$$\beta_{abc} = -\mathrm{i}\nabla_a\bar{\lambda}_{C'}\nabla_b\lambda_C \,, \quad \boldsymbol{\beta} = \beta_{[abc]} \,. \tag{18.94}$$

That is, $\boldsymbol{\alpha}$ has second derivatives of the spinor λ_A and $\boldsymbol{\beta}$ has squares of first derivatives of λ_A.

We first compute $\boldsymbol{\alpha}$. Observe that there is a commutator of covariant derivatives and hence we can replace it by the curvature tensor. However, what is surprising is that precisely the Einstein tensor appears. To see this, it is easier to work with the dual of $\boldsymbol{\alpha}$ defined by

$$^*\boldsymbol{\alpha} = \frac{1}{3!}\epsilon_{abcd}\alpha^{abc} \,. \tag{18.95}$$

We use the commutator relations

$$2\nabla_{[a}\nabla_{b]}\lambda_C = -\epsilon_{A'B'}X_{ABC}{}^E\lambda_E - \epsilon_{A\,B}\Phi_{A'B'C}{}^E\lambda_E \,, \tag{18.96}$$

where X_{ABCD} and $\Phi_{A'B'CD}$ are the curvature spinors. These spinors are defined in terms of the Riemann tensor $R_{abcd} = R_{AA'BB'CC''DD''}$ by

$$X_{ABCD} = \frac{1}{4}R_{AX'B}{}^{X'}{}_{CY'D}{}^{Y'} \,,$$
$$\Phi_{ABC'D'} = \frac{1}{4}R_{AX'B}{}^{X'}{}_{YC'}{}^{Y}{}_{D'} \,. \tag{18.97}$$

See [18.33] for further details on the curvature spinors. The Einstein tensor is given by

$$G_{ab} = -6\Lambda g_{ab} - \Phi_{ab} \,, \tag{18.98}$$

where Λ is given by

$$\Lambda = \frac{1}{6}X_{AB}{}^{AB} \,. \tag{18.99}$$

We also use the identities

$$X_{ABC}{}^B = 3\Lambda\epsilon_{AC} \,, \tag{18.100}$$

and

$$\epsilon_{abcd} = \mathrm{i}\left(\epsilon_{AC}\epsilon_{BD}\epsilon_{A'D'}\epsilon_{B'C'} - \epsilon_{AD}\epsilon_{BC}\epsilon_{A'C'}\epsilon_{B'D'}\right) . \tag{18.101}$$

Then we obtain

$$^*\boldsymbol{\alpha} = -\frac{1}{2\cdot 3!}\xi_e G^{ef} \,, \tag{18.102}$$

and hence

$$\boldsymbol{\alpha} = -\frac{1}{2\cdot 3!}\xi_e G^{ef}\epsilon_{fabc} \,. \tag{18.103}$$

The expressions (18.102) and (18.103) are pure tensorial expressions.

To compute $\boldsymbol{\beta}$ we proceed in a similar form. We work first with the dual

$$^*\boldsymbol{\beta} = \frac{1}{3!}\epsilon_{abcd}\beta^{bcd} \,. \tag{18.104}$$

It is important (we will see why later) to split the covariant derivative ∇_a into its temporal and spatial component. Let t^a denote the unit time-like normal to the surface S and h_{ab} be the intrinsic metric of the surface. We define the spatial $\mathcal{D}_a$ derivative as

$$\mathcal{D}_a = h_a{}^b\nabla_b \,. \tag{18.105}$$

Note that $\mathcal{D}_a$ is not the covariant derivative D of the intrinsic metric h used in the previous sections; they are related by the equation

$$\mathcal{D}_{AB}\lambda_C = D_{AB}\lambda_C + \frac{1}{\sqrt{2}}\pi_{ABC}{}^D\lambda_D \,, \tag{18.106}$$

where $\pi_{ABCD} = \pi_{(AB)(CD)}$ is the spinor representation of the second fundamental form of the surface.

From equation (18.105) we obtain

$$\nabla_a = \mathcal{D}_a - t_a t^b\nabla_b \,. \tag{18.107}$$

We replace the derivative ∇_a by (18.107) in the definition of $\boldsymbol{\beta}$ given by (18.94) and we compute the dual defined by (18.104) to obtain

$$^*\boldsymbol{\beta} = -\mathrm{i}\frac{1}{3!}\epsilon_{abcd}\left(\mathcal{D}^b\bar{\lambda}^{C'}\mathcal{D}^d\lambda^C\right) + W_a \,, \tag{18.108}$$

where

$$W_a = i\frac{1}{3!}\epsilon_{abcd}\left(t^b t^f \nabla_f \bar{\lambda}^{C'} \mathcal{D}^d \lambda^C + t^d t^f \nabla_f \lambda^C \mathcal{D}^b \bar{\lambda}^{C'}\right). \tag{18.109}$$

Note that W_a satisfies

$$t^a W_a = 0 . \tag{18.110}$$

Using the identity (18.101) we further decompose the first term on the right-hand side of (18.108)

$$\begin{aligned} &-\mathrm{i}\epsilon_{abcd}\mathcal{D}^b\bar{\lambda}^{C'}\mathcal{D}^d\lambda^C \\ &\quad= \mathcal{D}^b\bar{\lambda}^{B'}\mathcal{D}_{BA'}\lambda_A - \mathcal{D}^b\bar{\lambda}_{A'}\mathcal{D}_{AB'}\lambda_B , && (18.111)\\ &\quad= \mathcal{D}_{C'B}\bar{\lambda}^{C'}\mathcal{D}^B{}_{A'}\lambda_A + \mathcal{D}_{CB'}\lambda^C\mathcal{D}^{B'}{}_A\bar{\lambda}_{A'} \\ &\qquad - \mathcal{D}_b\lambda_A\mathcal{D}^b\bar{\lambda}_{A'} , && (18.112)\end{aligned}$$

where in the second line we have used the spinorial identity

$$\epsilon_{AB}\epsilon_{CD} + \epsilon_{BC}\epsilon_{AD} + \epsilon_{CA}\epsilon_{BD} = 0 . \tag{18.113}$$

Combining (18.108) and (18.111), we finally obtain

$$\begin{aligned} {}^*\boldsymbol{\beta} &= \mathcal{D}_{C'B}\bar{\lambda}^{C'}\mathcal{D}^B{}_{A'}\lambda_A + \mathcal{D}_{CB'}\lambda^C\mathcal{D}^{B'}{}_A\bar{\lambda}_{A'} \\ &\quad - \mathcal{D}_b\lambda_A\mathcal{D}^b\bar{\lambda}_{A'} + W_a . \end{aligned} \tag{18.114}$$

We are now in the position to perform the integral over S of $\mathrm{d}\boldsymbol{\Omega}$. Using (18.92), (18.103), and (18.114) we obtain

$$\begin{aligned} \int_S \mathrm{d}\boldsymbol{\Omega} = \int_S &\Big(4\pi T_{ab}\xi^b + \mathcal{D}_{C'B}\bar{\lambda}^{C'}\mathcal{D}^B{}_{A'}\lambda_A \\ &+ \mathcal{D}_{CB'}\lambda^C\mathcal{D}^{B'}{}_A\bar{\lambda}_{A'} - \mathcal{D}_b\lambda_A\mathcal{D}^b\bar{\lambda}_{A'}\Big) \\ &\times t^a \,\mathrm{d}v , \end{aligned} \tag{18.115}$$

where we have used Einstein equations

$$G_{ab} = 8\pi T_{ab} , \tag{18.116}$$

to replace the Einstein tensor by the energy-momentum tensor in the expression (18.103) for $\boldsymbol{\alpha}$. Note that the term W_a in (18.114) does not appear in the integral because it is orthogonal to t^a (18.110).

Assuming that the spinor λ^A has the fall-off behavior (18.83), then the identity (18.84) holds; using Stoke's theorem (18.81) (note that by (18.83) all other boundary integrals vanish) we finally obtain the famous Witten identity

$$\begin{aligned} P_a\overset{\circ}{\xi}{}^a = \frac{1}{8\pi}\int_S &\Big(4\pi T_{ab}\xi^b + \mathcal{D}_{C'B}\bar{\lambda}^{C'}\mathcal{D}^B{}_{A'}\lambda_A \\ &+ \mathcal{D}_{CB'}\lambda^C\mathcal{D}^{B'}{}_A\bar{\lambda}_{A'} - \mathcal{D}_b\lambda_A\mathcal{D}^b\bar{\lambda}_{A'}\Big) \\ &\times t^a \,\mathrm{d}v . \end{aligned} \tag{18.117}$$

If we assume that the energy-momentum tensor T_{ab} satisfies the dominant energy condition, then we have

$$T_{ab}\xi^a t^b \geq 0 , \tag{18.118}$$

and hence the first term in the integrand of (18.117) is non-negative. The last term in (18.117) is also non-negative since it involves the contraction with the Riemannian metric (which is negative definite) and the time-like vector $t^{AA'}$. To handle the second and third terms we impose the following equation on λ^A, which is called the Sen–Witten equation, [18.24, 34]

$$\mathcal{D}_{AB}\lambda^A = 0 . \tag{18.119}$$

Let us assume for the moment that there is a solution to this equation with the fall-off behavior (18.83). Then, from (18.117) we obtain

$$P_a\overset{\circ}{\xi}{}^a \geq 0 . \tag{18.120}$$

However, the constant null vector $\overset{\circ}{\xi}{}^a$ is arbitrary, and hence it follows that P_a should satisfy (18.8). To prove the rigidity part of theorem 18.1 the key ingredient is that $E = 0$ implies, again by the identity (18.117), that the spinor satisfies the equation

$$\mathcal{D}_{AB}\lambda_C = 0 , \tag{18.121}$$

that is, it is a covariant constant in the whole manifold. From this equation it can be deduced that the initial data on the surface correspond to the Minkowski spacetime (see [18.35] for the details of this argument).

It remains to discuss the solutions of equation (18.119). The existence of a solution to this equation under the required fall-off conditions (18.83) was proved in [18.30, 35, 36]. The main point is that equa-

tion (18.119) constitutes an elliptic system of first order for the two complex components of the spinor (this can be easily seen using the standard definition of ellipticity for systems, see, for example, [18.37] where this specific example is discussed). Hence this equation can, essentially, be handled as a Poisson equation. Solutions under weak decay conditions on the data of equation (18.119) was proved in [18.32].

Finally, let us discuss the proof of Theorem 18.2. This was done in [18.30, 38]. Remarkably, the proof is very similar, the only extra ingredient is that in Stoke's theorem we need to include an extra internal boundary term. This term has the form [18.30]

$$\int_{\partial B} \mathbf{\Omega} = \int_{\partial B} \left(\Theta_+ \lambda^0 \bar{\lambda}^{0'} - \rho' \lambda^1 \bar{\lambda}^{1'} + \lambda^{1'} \eth \lambda^0 - \bar{\lambda}^0 \bar{\eth} \lambda^1 \right) \mathrm{d}s\,. \quad (18.122)$$

In this equation Θ_+ is the null expansion defined previously by (18.53). The coefficient ρ' represents the ingoing null expansion on the surface, it is not important for our purposes. The functions λ^0 and λ^1 are the components of λ^A in an appropriated spinorial diad adapted to the two-surface ∂B. Finally, $\eth$ is a tangential differential operator to the two-surface. It can be shown that the appropriate inner Dirichlet boundary for equation (18.119) is to prescribe one of the components λ^0 or λ^1 (but not both) (this is a consequence of the elliptic character of this equation; see, for example [18.37] for an elementary treatment of this). If we prescribe $\lambda^1 = 0$ on ∂B and use that, by hypothesis this surface satisfies $\Theta_+ = 0$, then the boundary term (18.122) vanishes and we can proceed in the same way as above to prove the positivity of the energy. Note that without the condition $\Theta_+ = 0$ it is not possible to make the boundary term zero.

18.5 Further Results and Open Problems

In this article we have discussed only the positive energy theorem in three space dimensions. The spinorial proof presented in Sect. 18.4 works in any dimension [18.35], however in higher dimensions the existence of a spin structure involves restrictions on the topology of the manifold S. In Sect. 18.4 we used Weyl spinors, which are well adapted to four spacetime dimensions. For higher dimensions Dirac spinors are usually used. Other currently available proofs [18.23, 25] do not work in arbitrary high dimensions. To prove the positive energy theorem in all dimensions is one of the relevant open problems in this area.

The positive energy theorem can be refined to incorporate other physically relevant parameters. For example, using a similar argument as in Witten's proof it is possible to prove [18.38, 39] that the total mass M satisfies

$$M \geq |q|\,, \quad (18.123)$$

where q is the electric charge and the non-electromagnetic part of the energy momentum tensor must satisfy the appropriate conditions.

Recently, for axially symmetric black holes the following inequality was proved

$$M \geq \sqrt{|J|}\,, \quad (18.124)$$

Here J is the angular momentum of the black hole (see the review article [18.11] and references therein). The equality in (18.124) is achieved only for the extreme Kerr black hole. This inequality is proved for one black hole; a relevant open problem is to prove it for multiple black holes.

Another important extension is the Penrose inequality for black holes. The Riemannian black hole positivity theorem 18.2 can be generalized to include the area of the minimal surface, namely,

$$M \geq \sqrt{\frac{A}{16\pi}}\,, \quad (18.125)$$

with equality only for the Schwarzschild black hole. This result was proved by [18.25, 40]. The general case remains open, see the review article [18.41].

Finally, we have discussed the concept of total energy and linear momentum of an isolated system. It would be very desirable to have a quantity that measures the energy of a finite region of the spacetime. These kinds of quantities are called quasi-local mass. For a comprehensive review on this important open problem, see [18.28]. The following related, pure quasi-local, inequality for axially symmetric black holes has recently been proved

$$A \geq 8\pi|J|\,, \quad (18.126)$$

where A is the area and J is the quasi-local angular momentum of the black hole (see the review article [18.11] and references therein). The equality in (18.126) is achieved if and only if the local geometry of the black hole is equal to the extreme Kerr black hole local geometry. For non-axially symmetric black holes it is difficult to define the quasi-local angular momentum J [18.28]. An important open problem is to generalize the inequality (18.126) for non-axially symmetric black holes (or to find suitable counterexamples).

References

18.1 A. Ashtekar, G.T. Horowitz: Energy-momentum of isolated systems cannot be null, Phys. Lett. A **89**(4), 181–184 (1982)
18.2 V.I. Denisov, V.O. Solov'ev: The energy determined in general relativity on the basis of the traditional Hamiltonian approach does not have physical meaning, Theor. Math. Phys. **56**, 832–841 (1983)
18.3 H.L. Bray, P.T. Chruściel: The Penrose inequality. In: *The Einstein Equations and the Large Scale Behavior of Gravitational Fields*, ed. by P.T. Chruściel, H. Friedrich (Birkhäuser, Basel 2004) pp. 39–70
18.4 R. Bartnik: The mass of an asymptotically flat manifold, Commun. Pure App. Math. **39**(5), 661–693 (1986)
18.5 P. Chruściel: Boundary conditions at spatial infinity from a Hamiltonian point of view. In: *Topological Properties and Global Structure of Space-Time (Erice, 1985)*, NATO Advanced Science Institutes Series B: Physics., Vol. 138, ed. by P. Bergmann, V. de Sabbata (Plenum, New York 1986) pp. 49–59
18.6 P. T. Chruściel: Lectures on energy in General Relativity (2012), available online at http://homepage.univie.ac.at/piotr.chrusciel
18.7 J.M. Lee, T.H. Parker: The Yamabe problem, Bull. Am. Math. Soc. **17**(1), 37–91 (1987)
18.8 G. B. Cook: Initial data for numerical Relativity, Living Rev. Relativ. 3 (2000) 5, available online at http://www.livingreviews.org/Articles/Volume3/2000-5cook/
18.9 M. Cantor, D. Brill: The Laplacian on asymptotically flat manifolds and the specification of scalar curvature, Compositio Math. **43**(3), 317–330 (1981)
18.10 D. Brill: On the positive definite mass of the Bondi-Weber-Wheeler time-symmetric gravitational waves, Annu. Phys. **7**, 466–483 (1959)
18.11 S. Dain: Geometric inequalities for axially symmetric black holes, Class. Quantum Gravity **29**(7), 073001 (2012)
18.12 P. T. Chruściel: Lectures on mathemacital Relativity (2008), available online at http://homepage.univie.ac.at/piotr.chrusciel/papers/BeijingAll.pdf
18.13 H. L. Bray, J. L. Jauregui: A geometric theory of zero area singularities in general relativity (2009) 0909.0522
18.14 S. Dain, M.E. Gabach Clément: Extreme Bowen-York initial data, Class. Quantum Gravity **26**, 035020 (2009)
18.15 D.R. Brill, R.W. Lindquist: Interaction energy in geometrostatics, Phys. Rev. **131**, 471–476 (1963)
18.16 M. Alcubierre: *Introduction to 3+1 Numerical Relativity*, International Series of Monographs on Physics, Vol. 140 (Oxford Univ. Press, Oxford 2008)
18.17 C.W. Misner: Wormhole initial conditions, Phys. Rev. **118**, 1110–1111 (1960)
18.18 R.M. Wald: *General Relativity* (Univ. Chicago Press, Chicago 1984)
18.19 R. Osserman: *A survey of minimal surfaces*, 2nd edn. (Dover, New York 1986)
18.20 K. Martel, E. Poisson: Regular coordinate systems for Schwarzschild and other spherical space-times, Am. J. Phys. **69**, 476–480 (2001)
18.21 I. de Gentile Austria: *Superficies maximales con momento lineal en Schwarzschild*, Master's Thesis (Facultad de Matemática Astronomía y Física, Universidad Nacional de Córdoba 2010), available online at http://www.famaf.unc.edu.ar/dain/ivan-tf.pdf
18.22 D.R. Brill, P.S. Jang: The positive mass conjecture. In: *General Relativity and Gravitation*, Vol. 1, ed. by A. Held (Plenum, New York 1980) pp. 173–193
18.23 R. Schoen, S.T. Yau: On the proof of the positive mass conjecture in general relativity, Comm. Math. Phys. **65**(1), 45–76 (1979)
18.24 E. Witten: A new proof of the positive energy theorem, Commun. Math. Phys. **80**, 381–402 (1981)
18.25 G. Huisken, T. Ilmanen: The inverse mean curvature flow and the Riemannian Penrose inequality, J. Differ. Geom. **59**, 352–437 (2001)
18.26 R. Geroch: Energy extraccion, Ann. New York Acad. Sci. **224**, 108–117 (1973)
18.27 R. Penrose, W. Rindler: *Spinors and Space-Time*, Vol. 2 (Cambridge Univ. Press, Cambridge, 1986)
18.28 L.B. Szabados: Quasi-local energy-momentum and angular momentum in GR: A review article, Living Rev. Relativ. **7**, 4 (2004)
18.29 G.T. Horowitz, P. Tod: A relation between local and total energy in general relativity, Commun. Math. Phys. **85**, 429–447 (1982)
18.30 O. Reula, K.P. Tod: Positivity of the Bondi energy, J. Math. Phys. **25**(4), 1004–1008 (1984)
18.31 J.A. Nester: A New gravitational energy expression with a simple positivity proof, Phys. Lett. A **83**, 241 (1981)
18.32 P. Bizon, E. Malec: On Witten's positive energy proof for weakly asymptotically flat space-times, Class. Quantum Gravity **3**, L123 (1986)

18.33 R. Penrose, W. Rindler: *Spinors and Space-Time*, Vol. 1 (Cambridge Univ. Press, Cambridge, 1984)
18.34 A. Sen: On the existence of neutrino "zero-modes" in vacuum spacetimes, J. Math. Phys. **22**(8), 1781–1786 (1981)
18.35 T. Parker, C.H. Taubes: On Witten's proof of the positive energy theorem, Commun. Math. Phys. **84**(2), 223–238 (1982)
18.36 O. Reula: Existence theorem for solutions of Witten's equation and nonnegativity of total mass, J. Math. Phys. **23**(5), 810–814 (1982)
18.37 S. Dain: Elliptic systems. In: *Analytical and Numerical Approaches to Mathematical Relativity*, Lecture Notes in Physics, Vol. 692, ed. by J. Frauendiener, D. Giulini, V. Perlick (Springer, Berlin Heidelberg 2006) pp. 117–139
18.38 G.W. Gibbons, S.W. Hawking, G.T. Horowitz, M.J. Perry: Positive mass theorems for black holes, Commun. Math. Phys. **88**, 295–308 (1983)
18.39 G.W. Gibbons, C.M. Hull: A Bogomolny bound for general relativity and solitons in $N = 2$ supergravity, Phys. Lett. B **109**(3), 190–194 (1982)
18.40 H.L. Bray: Proof of the riemannian penrose conjecture using the positive mass theorem, J. Differ. Geom. **59**, 177–267 (2001)
18.41 M. Mars: Present status of the Penrose inequality, Class. Quantum Gravity **26**, 193001 (2009)

19. Conserved Charges in Asymptotically (Locally) AdS Spacetimes

Sebastian Fischetti, William Kelly, Donald Marolf

When a physical system is complicated and nonlinear, global symmetries and the associated conserved quantities provide some of the most powerful analytic tools to understand its behavior. This is as true in theories with a dynamical spacetime metric as for systems defined on a fixed spacetime background. Chapter 17 has already discussed the so-called Arnowitt–Deser–Misner (ADM) conserved quantities for asymptotically flat dynamical spacetimes, exploring in detail certain subtleties related to diffeomorphism invariance. In particular, it showed that the correct notion of global symmetry is given by the so-called asymptotic symmetries; equivalence classes of diffeomorphisms with the same asymptotic behavior at infinity. It was also noted that the notion of asymptotic symmetry depends critically on the choice of boundary conditions. Indeed, it is the imposition of boundary conditions that causes the true gauge symmetries to be only a subset of the full diffeomorphism group and thus allows the existence of nontrivial asymptotic symmetries at all.

This chapter will explore the asymptotic symmetries and corresponding conserved charges of asymptotically anti-de Sitter (AdS) spacetimes (and of the more general asymptotically locally AdS spacetimes). There are three excellent reasons for doing so. The first is simply to gain further insight into asymptotic charges in gravity by investigating a new example. Since empty AdS space is a maximally symmetric solution, asymptotically AdS spacetimes are a natural and simple choice. The second is that the structure one finds in the AdS context is actually much richer than that in asymptotically flat space. At the physical level, this point is deeply connected to the fact (see, e.g., [19.1]) that all multipole moments of a given field in AdS space decay at the same rate at infinity. So while in asymptotically flat space the far field is dominated mostly by monopole terms (with only subleading corrections from dipoles and higher multipoles) all terms contribute equally in AdS. It is therefore useful to describe not just global charges (e.g., the total energy) but also the local densities of these charges along the AdS boundary. In fact, it is natural to discuss an entire so-called *boundary stress tensor* $T^{ij}_{\rm bndy}$ rather than just the conserved charges it defines. For this reason, we take a somewhat different path to the construction of conserved AdS charges than was followed in Chap. 17. In particular, we will use covariant as opposed to Hamiltonian methods below, although we will show in Sect. 19.3 that the end results for conserved charges are equivalent.

The third reason to study conserved charges in AdS is their fundamental relation to the AdS/CFT correspondence [19.2–4], which may well be the most common application of general relativity in twenty-first century physics. While this is not the place for a detailed treatment of either string theory or AdS/CFT, no *Handbook of Spacetime* would be complete without presenting at least a brief overview of the correspondence. It turns out that this is easy to do once we have become familiar with $T^{ij}_{\rm bndy}$ and its cousins associated with other (nonmetric) fields. So at the end of this chapter (Sect. 19.4) we will take the opportunity to do so. We will introduce AdS/CFT from the gravity side without using tools from either string theory or CFT.

We will focus on such modern applications below, along with open questions. We make no effort to be either comprehensive or historical. Nevertheless, the reader should be aware that conserved charges for asymptotically AdS spacetimes were first constructed in [19.5], where the associated energy was also argued to be positive definite.

The plan for this chapter is as follows. After defining and discussing AdS asymptotics in Sect. 19.1, we construct variational principles for asymptotically AdS spacetimes in Sect. 19.2. This

allows us to introduce the boundary stress tensor $T^{ij}_{\rm bndy}$ and a similar so-called response function $\Phi_{\rm bndy}$ for a bulk scalar field. The conserved charges $Q[\xi]$ constructed from $T^{ij}_{\rm bndy}$ are discussed in Sect. 19.2.4, and we comment briefly on positivity of the energy in Sect. 19.2.5.

Section 19.3 then provides a general proof that the $Q[\xi]$ do indeed generate canonical transformations corresponding to the desired asymptotic symmetries. As a result, they agree (up to a possible choice of zero-point) with corresponding ADM-like charges $H[\xi]$ that would be constructed via the AdS-analogs of the Hamiltonian techniques used in Chap. 17. The interested reader can find such a Hamiltonian treatment in [19.6–8]. Below, we generally consider AdS gravity coupled to a simple scalar matter field. More complete treatments allowing more general matter fields can be found in e.g., [19.9–11]. Section 19.4 then defines the algebra $\mathcal{A}_{\rm bndy}$ of boundary observables and provides the above-mentioned brief introduction to AdS/CFT.

19.1 Asymptotically Locally AdS Spacetimes

This section discusses the notion of asymptotically locally AdS spacetimes. We begin by introducing the empty AdS space itself in Sect. 19.1.1 as a maximally-symmetric solution to the Einstein equations. We then explore the asymptotic structure of AdS, and in particular its conformal boundary. This structure is used to define the notions of asymptotically AdS (AAdS) and asymptotically locally AdS (AlAdS) spacetimes in Sect. 19.1.3. Section 19.1.4 then discusses the associated Fefferman–Graham expansion, which provides an even more detailed description of the asymptotics and which will play a critical role in constructing variational principles, the boundary stress tensor, and so forth in the rest of this chapter. Finally, Sect. 19.1.5 describes how the above structures transform under diffeomorphisms and introduces the notion of an asymptotic Killing vector field.

19.1.1 Anti-de Sitter Space

Let us begin with a simple geometric description of $(d+1)$-dimensional AdS space (AdS_{d+1}) building on the reader's natural intuition for flat geometries. We will, however, need to begin with a flat spacetime $\mathbb{M}^{2,d}$ of signature $(2,d)$ having two time-directions and d spatial directions, so that in natural coordinates $T^1, T^2, X^1, \ldots, X^d$ the line element takes the form

$$\mathrm{d}s^2 = -(\mathrm{d}T^1)^2 - (\mathrm{d}T^2)^2 + (\mathrm{d}X^1)^2 + \cdots + (\mathrm{d}X^d)^2 . \tag{19.1}$$

Consider the $(d+1)$-dimensional hyperboloid $\mathcal{H}$ of events in $\mathbb{M}^{2,d}$ satisfying

$$(T^1)^2 + (T^2)^2 - \sum_{i=1}^{d} (X^i)^2 = \ell^2 , \tag{19.2}$$

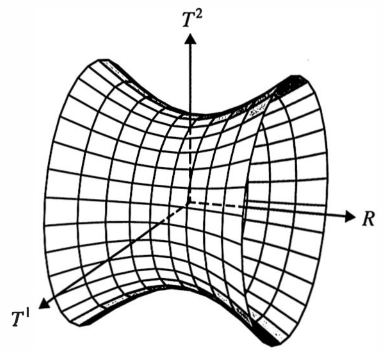


Fig. 19.1 The hyperboloid (19.2) embedded in $\mathbb{M}^{2,d}$, defining AdS space

which thus lie at a proper distance ℓ from the origin; see Fig. 19.1. This hyperboloid is sometimes known as the $d+1$ AdS space AdS_{d+1}, although we will follow a more modern tradition and save this name for a closely related (but much improved) spacetime that we have yet to introduce.

The isometries of $\mathcal{H}$ are given by symmetries of $\mathbb{M}^{2,d}$ preserved by (19.2). Such isometries form the group $SO(d,2)$, generated by the rotation in the T^1, T^2 plane together with two copies of the Lorentz group $SO(d,1)$ that act separately on $T^1, X^1, \ldots, X^d$ and $T^2, X^1, \ldots X^d$. This gives $(d+1)(d+2)/2$ independent symmetries so that $\mathcal{H}$ is maximally symmetric.

A simple way to parametrize the hyperboloid is to write $T^1 = \sqrt{\ell^2+R^2}\cos(\tau/\ell)$ and $T^2 = \sqrt{\ell^2+R^2}\sin(\tau/\ell)$, with $R^2 = \sum (X^i)^2$, so that the induced line element on $\mathcal{H}$ becomes

$$\mathrm{d}s^2_{\mathrm{AdS}_{d+1}} = -\left(\frac{R^2}{\ell^2}+1\right)\mathrm{d}\tau^2 + \frac{\mathrm{d}R^2}{\frac{R^2}{\ell^2}+1} + R^2\,\mathrm{d}\Omega^2_{d-1}\,. \tag{19.3}$$

On $\mathcal{H}$, the coordinate τ is periodic with period 2π. But this makes manifest that $\mathcal{H}$ contains closed time-like curves such as, for example, the worldline $R=0$. It is thus useful to unwrap this time direction by passing to the universal covering space of $\mathcal{H}$ or, more concretely, by removing the periodic identification of τ (so that τ now lives on $\mathbb{R}$ instead of S^1). We will refer to this covering space as the AdS space AdS_{d+1} with scale ℓ. Of course, the line element remains that of (19.3). Since any Killing field of $\mathcal{H}$ lifts readily to the covering space, AdS_{d+1} remains maximally symmetric with isometry group given by (a covering group of) $SO(d,2)$.

The coordinates used in (19.3) are called global coordinates, since they cover all of AdS. We can introduce another useful set of coordinates, called Poincaré coordinates, by setting $z=\ell^2/(T^1+X^d)$, $t=\ell T^2/(T^1+X^d)$, and $x^i = \ell X^i/(T^1+X^d)$ for $i=1,\ldots,d-1$. The metric then becomes

$$\mathrm{d}s^2_{\mathrm{AdS}_{d+1}} = \frac{\ell^2}{z^2}\left(-\mathrm{d}t^2 + \sum_{i=1}^{d-1}\left(\mathrm{d}x^i\right)^2 + \mathrm{d}z^2\right). \tag{19.4}$$

Poincaré coordinates take their name from the fact that they make manifest a (lower dimensional) Poincaré symmetry associated with the d coordinates t, x^i. As is clear from their definitions, these coordinates cover only the region of AdS where $T^1+X^d>0$. This region is called the Poincaré patch. While we will not make significant use of (19.4) below, we mention these coordinates here since they arise naturally in many discussions of AdS/CFT which the reader may encounter in the future.

Since AdS is maximally symmetric, its Riemann tensor can be written as an appropriately symmetrized combination of metric tensors

$$R_{\mu\nu\sigma\lambda} = \frac{1}{\mathrm{d}(d+1)}R(g_{\mu\sigma}g_{\nu\lambda} - g_{\mu\lambda}g_{\nu\sigma})\,. \tag{19.5}$$

A computation shows that the scalar curvature of AdS is $R = -\mathrm{d}(d+1)/\ell^2$ and thus that AdS solves the vacuum Einstein field equations with cosmological constant $\Lambda = -\mathrm{d}(d-1)/2\ell^2$

$$R_{\mu\nu} - \tfrac{1}{2}Rg_{\mu\nu} + \Lambda g_{\mu\nu} = 0\,. \tag{19.6}$$

In this sense, AdS is a generalization of flat space to $\Lambda<0$.

19.1.2 Conformal Structure and Asymptotic Symmetries of AdS

We now turn to the asymptotic structure of AdS, which was seen in Chap. 17 to be a crucial ingredient in the construction of conserved charges. It is useful to introduce a new radial coordinate $r_* = \arctan(R/\ell)$, so that the line element becomes

$$\mathrm{d}s^2_{\mathrm{AdS}_{d+1}} = \frac{\ell^2}{\cos^2(r_*)} \times\left[-\frac{\mathrm{d}\tau^2}{\ell^2} + \mathrm{d}r_*^2 + \sin^2(r_*)\,\mathrm{d}\Omega^2_{d-1}\right]. \tag{19.7}$$

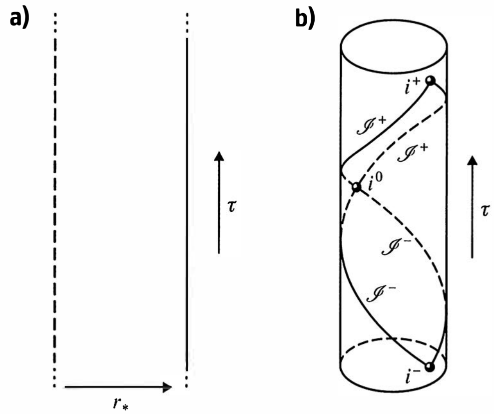


Fig.19.2a,b Conformal diagrams of AdS_{d+1}, showing both the global spacetime and the region covered by the Poincaré patch. In both figures, the τ direction extends infinitely to the future and to the past. In (**a**), a full S^{d-1} of symmetry has been suppressed, leaving only the τ, r_* coordinates of (19.7). The *dotted line* corresponds to $r_* = 0$. In (**b**), one of the angular directions has been shown explicitly to guide the reader's intuition; the axis of the cylinder corresponds to the *dotted line* in (**a**). The Poincaré patch covers a *wedge-shaped region* of the interior of the cylinder which meets the boundary at the *lines* marked $\mathscr{I}^\pm$ and the points marked $i^\pm, i^0$. These loci form the null, time-like, and space-like infinities of the associated region (conformal to Minkowski space) on the AdS boundary

We can immediately identify $r_* = \pi/2$ as a conformal boundary, leading to the conformal diagrams shown in Fig. 19.2. (For readers not familiar with such diagrams, Chap. 25 will give a brief introduction.)

It is evident from the conformal diagram that AdS is not globally hyperbolic. In order to evolve initial data on some space-like surface Σ arbitrarily far forward (or backward) in time, one needs to supply additional information in the form of boundary conditions at the conformal boundary. Such boundary conditions will be discussed in detail in Sect. 19.2, where they will play critical roles in our discussion of conserved charged.

Although the line element (19.7) diverges at $r_* = \pi/2$, the rescaled metric

$$\hat{g} = \frac{\cos^2(r_*)}{\ell^2} g_{AdS_{d+1}} \tag{19.8}$$

defines a smooth manifold with boundary. In particular, the metric induced by $\hat{g}$ at $r_* = \pi/2$ is just that of the flat cylinder $\mathbb{R} \times S^{d-1}$, also known as the Einstein static universe (ESU). The manifold with boundary will be called M and the boundary itself (at $r_* = \pi/2$) will be called ∂M. Of course, we could equally well have considered the more general rescaled metric

$$\hat{g}' = \frac{\cos^2(r_*)}{\ell^2} e^{2\sigma} g_{AdS_{d+1}} , \tag{19.9}$$

where σ is an arbitrary smooth function on M. This metric is also nonsingular at $r_* = \pi/2$, but the induced geometry on ∂M is now only conformal to $\mathbb{R} \times S^{d-1}$. The choice of a particular rescaled metric (19.9) (or, equivalently, of a particular rescaling factor $\frac{\cos^2(r_*)}{\ell^2} e^{2\sigma}$) determines a representative of the corresponding conformal class of boundary metrics. This choice (which still allows great freedom to choose σ away from ∂M) is known as the choice of conformal frame. We shall often call this representative *the boundary metric,* where it is understood that the above choices must be made for this term to be well defined.

Although it is not critical for our discussion below, the reader should be aware of the asymptotic structure of the Poincaré patch and how it relates to that of global AdS as discussed above. From (19.4) we see that the conformal boundary lies at $z = 0$. The rescaled metric

$$\hat{g} = \frac{z^2}{\ell^2} g_{AdS_{d+1}} \tag{19.10}$$

is regular at $z = 0$, where the induced metric is just d-dimensional Minkowski space. Now, it is well known [19.12] that Minkowski space $\mathbb{M}^{1,d-1}$ is conformally equivalent to a patch of the Einstein static universe $\mathbb{R} \times S^{d-1}$. We conclude that $z = 0$ of the Poincaré patch is a diamond-shaped piece of ∂M, as shown on the right-hand side in Fig. 19.2.

In the interior of AdS the Poincaré patch covers a wedge-shaped region. This can be thought of as follows: future-directed null geodesics fired from i^- in Fig. 19.2 are focused onto i^0; these geodesics are generators of a null hypersurface, which we shall call the past Poincaré horizon $\mathcal{H}^-_{\text{Poincaré}}$. Likewise, future-directed null geodesics fired from i^0 are focused onto i^+, generating the future Poincaré horizon $\mathcal{H}^+_{\text{Poincaré}}$. The Poincaré patch of AdS is the wedge enclosed by these horizons.

19.1.3 A Definition of Asymptotically Locally AdS Spacetimes

As we saw in Chap. 17, when the spacetime metric is dynamical the choice of boundary conditions

plays an especially key role in constructions of conserved charges. In this chapter we consider boundary conditions which force the spacetime to behave asymptotically in a manner at least locally similar to (19.3). It turns out to be useful to proceed by using the notion of a conformally rescaled metric $\hat{g}$, which extends sufficiently smoothly to the boundary (see Chap. 25 for a further discussion of this technique). After imposing the equations of motion, this $\hat{g}$ will allow us to very quickly define both asymptotically AAdS and AlAdS. Below, we follow [19.9, 13–18].

To begin, recall that our discussion of pure AdS above made use of the fact that the unphysical metrics defined in (19.8) and (19.10) could be extended to the conformal boundary ∂M of AdS. We can generalize this notion by considering any manifold M (often called *the bulk*) with boundary ∂M and allowing metrics g which are singular on ∂M but for which but there exists a smooth function Ω satisfying $\Omega|_{\partial M} = 0$, $(\mathrm{d}\Omega)|_{\partial M} \neq 0$ (where $|_{\partial M}$ denotes the pull-back to ∂M), and $\Omega > 0$ on all of M, such that

$$\hat{g} = \Omega^2 g \tag{19.11}$$

can be extended to all of M as a sufficiently smooth non-degenerate metric for which the induced metric on ∂M has a Lorentz signature. We will discuss what is meant by sufficiently smooth in more detail in Sect. 19.1.4, but for the purposes of this section one may take $\hat{g}$ to be C^2 (so that its Riemann tensor is well defined). Note that $\hat{g}$ is not unique; given any allowed Ω one is always free to choose

$$\Omega' = e^{\sigma}\Omega\,, \tag{19.12}$$

for arbitrary smooth σ on M. Thus, as before, the notion of a particular boundary metric on ∂M is well defined only after one has chosen some conformal frame. However, the bulk metric g does induce a unique conformal structure on ∂M. The function Ω is termed the defining function of the conformal frame. The above structure is essentially that of Penrose's conformal compactifications [19.19], except that the Lorentz signature of ∂M forbids M to be fully compact. In particular, future and past infinity are not part of ∂M.

In vacuum Einstein–Hilbert gravity with cosmological constant (19.6), we define an asymptotically locally AdS spacetime to be a spacetime (g, M) as above that solves the Einstein equations (19.6). A key feature of this definition is that it makes no restriction on the conformal structure, or even the topology of the boundary, save that it be compatible with having a Lorentz signature metric. For an AlAdS spacetime to be what we will call AAdS, the induced boundary metric must be conformal to $\mathbb{R}\times S^{d-1}$. The reader should be aware that in the literature, the term *AAdS* is sometimes used synonymously with *AlAdS*. Here we emphasize the distinction between the two for pedagogical purposes, as only AAdS spacetimes can truly be said to approach global AdS near ∂M.

To show that AlAdS spacetimes do, in fact, approach (19.5) requires the use of the Einstein equations. By writing $g_{\mu\nu} = \Omega^{-2}\hat{g}_{\mu\nu}$, a straightforward calculation then shows [19.17] that near ∂M we have

$$\begin{aligned} R_{\mu\nu\sigma\lambda} = &-|\mathrm{d}\Omega|^2_{\hat{g}}\,(g_{\mu\sigma}g_{\nu\lambda} - g_{\nu\sigma}g_{\mu\lambda}) \\ &+ \mathcal{O}(\Omega^{-3})\,, \end{aligned} \tag{19.13}$$

where

$$|\mathrm{d}\Omega|^2_{\hat{g}} \equiv \hat{g}^{\mu\nu}\partial_\mu\Omega\partial_\nu\Omega \tag{19.14}$$

extends smoothly to ∂M. Note that since g has a second-order pole at ∂M, the leading-order term in (19.13) is of order Ω^{-4}. The Einstein field equations then imply that

$$|\mathrm{d}\Omega|^2_{\hat{g}} = \frac{1}{\ell^2} \quad \text{on } \partial M\,. \tag{19.15}$$

It follows that Riemann tensor (19.13) of an AlAdS spacetime near ∂M looks like that of pure AdS (19.5). Further details of the asymptotic structure (and of the approach to (19.3) for the AAdS case) are elucidated by the Fefferman–Graham expansion near ∂M, to which we now turn.

19.1.4 The Fefferman–Graham Expansion

The term asymptotically (locally) AdS suggests that the spacetime metric g should (locally) approach (19.3), at least with a suitable choice of coordinates. This is far from manifest in the definitions above, but it turns out to be a consequence of the Einstein equations. In fact, these equations imply that the asymptotic structure is described by a so-called Fefferman–Graham expansion [19.20].

The basic idea of this expansion is to first choose a convenient set of coordinates and then to attempt a power-series solution to the Einstein equations. Since the Einstein equations are second order, this leads to a second-order recursion relation for the coefficients of the power series. For, say, simple ordinary differential

equations, one would expect the free data in the power series to be parametrized by two of the coefficients. The structure that emerges from the Einstein equations is similar, except for the presence of constraint equations similar to those described in Chap. 17. As we will briefly describe below, the constraint equations lead to corresponding constraints on the two otherwise free coefficients. We continue to consider the vacuum case (19.6).

Let us begin by introducing the so-called Fefferman–Graham coordinates on some finite neighborhood U of ∂M. To do so, note that since the defining function Ω is not unique it is possible to choose a σ in (19.12) such that the modified defining function $z := \Omega'$ obeys

$$|\mathrm{d}z|^2_{\hat{g}} = \frac{1}{\ell^2} \tag{19.16}$$

on U, where $\hat{g} = z^2 g$. In fact, we can do so with $\sigma|_{\partial M} = 1$, so that we need not change the conformal frame. We can then take the defining function z to be a coordinate near the boundary; the notation z is standard for this so-called *Fefferman–Graham radial coordinate*. We choose the other coordinates x^i to be orthogonal to z in U (according to the metric $\hat{g}$). The metric in these so-called Fefferman–Graham coordinates will then take the form

$$\mathrm{d}s^2 = \frac{\ell^2}{z^2}\left(\mathrm{d}z^2 + \gamma_{ij}(x,z)\,\mathrm{d}x^i\,\mathrm{d}x^j\right), \tag{19.17}$$

where $i = 0, \ldots, d$. By construction, γ_{ij} can be extended to ∂M, so it should admit an expansion (at least to some order) in nonnegative powers of z

$$\gamma_{ij}(x,z) = \gamma^{(0)}_{ij}(x) + z\gamma^{(1)}_{ij}(x) + \cdots . \tag{19.18}$$

Note that $\gamma^{(0)}_{ij}$ defines the metric $\gamma^{(0)}$ on ∂M in this conformal frame.

Since the Einstein equations are second- order partial differential equations, plugging in the ansatz (19.18) leads to a second-order recursion relation for the $\gamma^{(n)}$. For odd d this recursion relation admits solutions for all $\gamma^{(n)}$. After specifying $\gamma^{(0)}$, one finds that all $\gamma^{(n)}$ with $n < d$ are uniquely determined (and, in fact $\gamma^{(n)}$ vanishes for all odd $n < d$). For example, for $d > 2$ one finds [19.18]

$$\gamma^{(2)}_{ij} = \frac{-1}{d-2}\left(\mathcal{R}_{ij} - \frac{1}{2(d-1)}\mathcal{R}\gamma^{(0)}_{ij}\right), \tag{19.19}$$

where $\mathcal{R}$, $\mathcal{R}_{ij}$ are, respectively, the Ricci tensor and Ricci scalar of $\gamma^{(0)}$.

However, new data enters in $\gamma^{(d)}$. This new data is subject to constraints that are analogous to those discussed in the Hamiltonian formalism in Chap. 17. Indeed, these constraints may be derived by considering the analogs of the Hamiltonian and momentum constraints on surfaces with $z = constant$. They determine the trace and divergence of $\gamma^{(d)}$ (again for d odd) through

$$(\gamma^{(0)})^{ij}\gamma^{(d)}_{ij} = 0\,, \quad (\gamma^{(0)})^{ki}D_k\gamma^{(d)}_{ij} = 0\,, \tag{19.20}$$

where D_k is the $\gamma^{(0)}$-compatible derivative operator on ∂M (where we think of all $\gamma^{(n)}$ as being defined). We will give a short argument for (19.20) in Sect. 19.2.4. Once we have chosen any $\gamma^{(d)}$ satisfying (19.20), the recursion relation can then be solved order-by-order to express all higher $\gamma^{(n)}$ in terms of $\gamma^{(0)}$ and $\gamma^{(d)}$. Of course, the series (19.17) describes only the asymptotic form of the metric. There is no guarantee that there is, in fact, a smooth solution in the interior matching this asymptotic data, or that such a smooth interior solution is unique when it exists.

The situation is slightly more complicated for even d, where the recursion relations for the ansatz (19.18) break down at the order at which $\gamma^{(d)}$ would appear. To proceed, one must allow logarithmic terms to arise at this order and use the more general ansatz

$$\begin{aligned}\gamma_{ij}(x,z) = \gamma^{(0)}_{ij} &+ z^2\gamma^{(2)}_{ij} + \cdots + z^d\gamma^{(d)}_{ij} \\ &+ z^d\bar{\gamma}^{(d)}_{ij}\log z^2 + \cdots\,,\end{aligned} \tag{19.21}$$

where, since the structure is identical for all d up to order $n = d$, we have made manifest that $\gamma^{(n)} = 0$ for all odd $n < d$. The higher-order terms represented by $\cdots$ include both higher even powers of z and such terms multiplied by $\log z$. One finds that $\bar{\gamma}^{(d)}$ is fully determined by $\gamma^{(0)}$ and satisfies

$$(\gamma^{(0)})^{ij}\bar{\gamma}^{(d)}_{ij} = 0\,, \quad (\gamma^{(0)})^{ki}D_k\bar{\gamma}^{(d)}_{ij} = 0\,. \tag{19.22}$$

For example, for $d = 2, 4$, one obtains [19.18]

$$\bar{\gamma}^{(2)}_{ij} = 0\,, \tag{19.23}$$

$$\begin{aligned}\bar{\gamma}^{(4)}_{ij} &= \tfrac{1}{8}\mathcal{R}_{ikjl}\mathcal{R}^{kl} - \tfrac{1}{48}D_iD_j\mathcal{R} \\ &+ \tfrac{1}{16}D^2\mathcal{R}_{ij} - \tfrac{1}{24}\mathcal{R}\mathcal{R}_{ij} \\ &+ \left(\tfrac{-1}{96}D^2\mathcal{R} + \tfrac{1}{96}\mathcal{R}^2 - \tfrac{1}{32}\mathcal{R}_{kl}\mathcal{R}^{kl}\right)\gamma^{(0)}_{ij}\,,\end{aligned} \tag{19.24}$$

where $\mathcal{R}_{ijkl}$ is the Riemann tensor of $\gamma^{(0)}$ and indices are raised and lowered with $\gamma^{(0)}$. However, $\gamma^{(d)}$ may again be chosen freely subject to dimension-dependent conditions that fix its divergence and trace. As examples, one finds [19.18]

$$d=2: \quad (\gamma^{(0)})^{ij}\gamma^{(d)}_{ij} = \tfrac{-1}{2}\mathcal{R}\,, \quad D^i\gamma^{(d)}_{ij} = \tfrac{-1}{2}D_j\mathcal{R}\,, \tag{19.25}$$

$$d=4: \quad (\gamma^{(0)})^{ij}\gamma^{(d)}_{ij} = \tfrac{1}{16}\left(\mathcal{R}_{ij}\mathcal{R}^{ij} - \frac{2}{9}\mathcal{R}^2\right), \tag{19.26}$$

$$D^i\gamma^{(d)}_{ij} = \tfrac{1}{8}\mathcal{R}_i{}^kD^i\mathcal{R}_{kj} - \tfrac{1}{32}D_j\left(\mathcal{R}^{ik}\mathcal{R}_{ik}\right) + \tfrac{1}{288}\mathcal{R}D_j\mathcal{R}\,. \tag{19.27}$$

The higher terms in the series are again uniquely determined by $\gamma^{(0)}$, $\gamma^{(d)}$.

In general, the terms $\gamma^{(n)}$ become more and more complicated at each order. However, the expansion simplifies when $\gamma^{(0)}_{ij}$ is conformally flat and $\gamma^{(d)}_{ij} = 0$. In this case, one finds [19.21] that the recursion relation can be solved exactly and terminates at order z^4. In particular, the bulk metric so obtained is also conformally flat and is thus locally AdS_{d+1}. For $d=2$, the Fefferman–Graham expansion can be integrated exactly for any $\gamma^{(0)}$, $\gamma^{(d)}$ and always terminates at order z^4 to define a metric that is locally AdS_3.

19.1.5 Diffeomorphisms and Symmetries in AlAdS

The reader of this Handbook is by now well aware of the important roles played by diffeomorphisms in understanding gravitational physics. Let us, therefore, pause briefly to understand how such transformations affect the structures defined thus far. We are interested in diffeomorphisms of our manifold M with boundary ∂M. By definition, any such diffeomorphism must map ∂M to itself; i. e., it also induces a diffeomorphism of ∂M. As usual in physics, we consider diffeomorphisms (of M) generated by vector fields ξ; the corresponding diffeomorphism of ∂M is generated by some $\hat{\xi}$, which is just the restriction of ξ to ∂M (where by the above it must be tangent to ∂M).

Of course, the metric g transforms as a tensor under this diffeomorphism. However, if we think of the diffeomorphism as acting only on dynamical variables of the theory then the defining function $z = \Omega$ does not transform at all, and in particular does not transform like a scalar field. This means that the rescaled metric $\hat{g} = z^2 g$ does *not* transform like a tensor, and neither does the boundary metric $\gamma^{(0)}$. Instead, the diffeomorphism induces an additional conformal transformation on ∂M, i. e., a change of conformal frame.

We can make this explicit by considering diffeomorphisms that preserve the Fefferman–Graham gauge conditions, i. e., which satisfy

$$\delta g_{zz} = 0 = \delta g_{iz} \tag{19.28}$$

for

$$\delta g_{\mu\nu} = \pounds_\xi g_{\mu\nu} = \nabla_\mu \xi_\nu + \nabla_\nu \xi_\mu\,, \tag{19.29}$$

where we use $\pounds_\xi$ to denote Lie derivatives along ξ and ∇_μ is the covariant derivative compatible with the metric g on M. Let us decompose the components $\delta g_{\mu\nu}$ into

$$\pounds_\xi g_{zz} = \frac{2\ell}{z}\partial_z\left(\frac{\ell}{z}\xi^z\right), \tag{19.30}$$

$$\pounds_\xi g_{iz} = \frac{\ell^2}{z^2}\left(\partial_i\xi^z + \gamma_{ij}\partial_z\xi^j\right), \tag{19.31}$$

$$\pounds_\xi g_{ij} = \frac{\ell^2}{z^2}\left(\pounds_{\hat{\xi}}\gamma_{ij} + z^2\partial_z\left(z^{-2}\gamma_{ij}\right)\xi^z\right), \tag{19.32}$$

where $\pounds_{\hat{\xi}}$ is the Lie derivative with respect to $\hat{\xi}$ on ∂M. These conditions can be integrated using (19.28) to obtain

$$\xi^z = z\hat{\xi}^z(x)\,, \tag{19.33}$$

$$\xi^i = \hat{\xi}^i(x) - \partial_j\hat{\xi}^z\int_0^z z'\gamma^{ji}(z')\,\mathrm{d}z'\,, \tag{19.34}$$

where $\hat{\xi}^z$ and $\hat{\xi}^i$ are an arbitrary function and vector field on ∂M (which we may transport to any $z = \textit{constant}$ surface by using the given coordinates to temporarily identify that surface with ∂M). In particular, for $\hat{\xi}^i = 0$ we find

$$g_{ij} + \delta g_{ij} = \frac{\ell^2}{z^2}(1 - 2\hat{\xi}^z)\gamma^{(0)}_{ij} + \mathcal{O}(z^0)\,. \tag{19.35}$$

Thus the boundary metric transforms as $\gamma^{(0)} \to e^{-2\hat{\xi}^z}\gamma^{(0)}_{ij}$. Such transformations are called conformal transformations by relativists and Weyl transformations by particle physicists; we will use the former, but the

reader will find both terms in various treatments of AlAdS spacetimes. This is precisely the change of conformal frame mentioned above.

Let us now turn to the notion of symmetry. As in Chap. 17, we might be interested either in an exact symmetry of some metric g, generated by a Killing vector field (KVF) satisfying $\nabla_{(\nu}\xi_{\mu)} = 0$, or in some notion of asymptotic symmetry. We will save the precise definition of an asymptotic symmetry for Sect. 19.2.3 as, strictly speaking, this first requires the construction an appropriate variational principle and a corresponding choice of boundary conditions. However, we will discuss the closely related (but entirely geometric) notion of an asymptotic Killing field below.

Suppose first that ξ is, indeed, a KVF of g so that $\pounds_\xi g = 0$. It is clear that there are two cases to consider. Either $\pounds_\xi \Omega = 0$ (in which case we say that ξ is compatible with Ω) or $\pounds_\xi \Omega \neq 0$ (in which case we say that ξ is not compatible with Ω). In the former case we clearly have $\pounds_\xi \hat{g} = \pounds_\xi(\Omega^2 g) = 0$ so that ξ is also a Killing field of $\hat{g}$. However, more generally we have seen that the corresponding diffeomorphism changes $\hat{g}$ by a conformal factor. The generators of such diffeomorphisms are called conformal Killing fields of $\hat{g}$ (see, e.g., [19.12, Appendix C.3]) and satisfy

$$\begin{aligned}\pounds_\xi \hat{g}_{\mu\nu} &= (\pounds_\xi \ln \Omega^2)\hat{g}_{\mu\nu} \Rightarrow 2\widehat{\nabla}_{(\mu}\xi_{\nu)} \\ &= \frac{2}{d+1}\left(\widehat{\nabla}_\sigma \xi^\sigma\right)\hat{g}_{\mu\nu}\,,\end{aligned} \tag{19.36}$$

where $\widehat{\nabla}$ is the covariant derivative compatible with $\hat{g}$ and indices on ξ^μ are lowered with $\hat{g}_{\mu\nu}$. Note that the induced vector field $\hat{\xi}$ on ∂M is again a conformal Killing field of $\gamma^{(0)}$.

This suggests that we define an asymptotic Killing field to be any vector field ξ that satisfies (19.36) to leading order in Ω at ∂M. If we ask that ξ also preserve Fefferman–Graham gauge we may then expand (19.33) and (19.34) and insert into (19.36) to obtain

$$\xi^z = z\hat{\xi}^z(x)\,, \tag{19.37}$$

$$\xi^i = \hat{\xi}^i(x) - \tfrac{1}{2}z^2(\gamma^{(0)})^{ij}\partial_j\hat{\xi}^z + \mathcal{O}(z^4)\,, \tag{19.38}$$

$$\pounds_{\hat{\xi}}\gamma^{(0)}_{ij} - \frac{2}{d+1}\left(D_k\hat{\xi}^k + \hat{\xi}^z\right)\gamma^{(0)}_{ij} = 0\,. \tag{19.39}$$

Taking the trace of the condition (19.39) shows that $\hat{\xi}^z = \frac{1}{d}D_i\hat{\xi}^i$, so (19.39) is the conformal Killing equation for $\hat{\xi}$ with respect to $\gamma^{(0)}$. In other words, conformal Killing fields $\hat{\xi}$ of $\gamma^{(0)}$ are in one-to-one correspondence with asymptotic Killing fields of g which preserve Fefferman–Graham gauge, where the equivalence relation is given by agreement to the order shown in (19.37).

19.1.6 Gravity with Matter

Our treatment above has focused on vacuum gravity. It is useful to generalize the discussion to include matter fields, both to see how this influences the above result and also to better elucidate the general structure of asymptotically AdS field theory. Indeed, readers new to dynamics in AdS space will gain further insight from Sect. 19.1.4 if they re-read it after studying the treatment of the free scalar field below. We use a single scalar as an illustrative example of matter fields; see [19.9, 10] for more general discussions.

For simplicity, we first consider a massive scalar field in a fixed AlAdS$_{d+1}$ gravitational background, which we take to be in Fefferman–Graham form (19.17). This set-up is often called the probe approximation as it neglects the back-reaction of the matter on the spacetime. The action is as usual

$$S^{\text{Bulk}}_\phi = -\tfrac{1}{2}\int \mathrm{d}^{d+1}x\sqrt{|g|}\left(g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi + m^2\phi^2\right). \tag{19.40}$$

We study the behavior of solutions near the boundary $z=0$ by seeking solutions which behave at leading order like z^Δ for some power Δ. The equation of motion

$$(-\Box + m^2)\phi = 0 \tag{19.41}$$

then requires $(m\ell)^2 = \Delta(\Delta - d)$, yielding two independent small-z behaviors $z^{\Delta_\pm}$. Here we have defined $\Delta_\pm = d/2 \pm \nu$, with $\nu \equiv \sqrt{(d/2)^2 + (m\ell)^2}$. A priori, it seems that we should consider only $\nu \geq \nu_{\min}$ for some $\nu_{\min} > 0$, since one might expect $(m\ell)^2 \geq 0$. However, it can be shown [19.22] that scalar fields with small tachyonic masses in AdS$_{d+1}$ are stable as long as the mass satisfies the so-called Breitenlohner–Freedman (BF) bound $(m\ell)^2 \geq -d^2/4 =: m^2_{\mathrm{BF}}$; we, therefore, consider $\nu \geq 0$. The essential points here are: i) It is only for $|(m\ell)^2| \gg 1$ that the flat-space approximation must hold, so for small $|(m\ell)^2|$ the behavior

can differ significantly from that of flat space; and ii) as noted above, the fact that AdS is not globally hyperbolic means that we must impose boundary conditions at ∂M. These boundary conditions generally require ϕ to vanish on ∂M. So even for $m^2 = 0$ we would exclude the *zero mode* $\phi =$ constant. For a given boundary condition, the spectrum of modes turns out to be discrete. As a result, we may lower m^2 a finite amount below zero before a true instability develops.

The asymptotic analysis above suggests that we seek a solution of the form

$$\phi(x,z) = z^{\Delta_-}(\phi^{(0)} + z^2\phi^{(2)} + \cdots) + z^{\Delta_+}(\phi^{(2\nu)} + z^2\phi^{(2\nu+2)} + \cdots)\,. \tag{19.42}$$

For noninteger ν the equation of motion can be solved order-by-order in z to uniquely express all coefficients in terms of $\phi^{(0)}$ and $\phi^{(2\nu)}$. However, for integer ν the difference $\Delta_+ - \Delta_-$ is an even integer and the two sets of terms in (19.42) overlap. This notational issue is connected to a physical one: keeping only even-integer powers of z (times z^{Δ_-}) does not allow enough freedom to solve the resulting recursion relation; there is no solution at order $d - 2\Delta_-$. To continue further we must introduce a logarithmic term and write

$$\phi(x,z) = z^{\Delta_-}(\phi^{(0)} + z^2\phi^{(2)} + \cdots) + z^{\Delta_+}\log z^2(\psi^{(2\nu)} + z^2\psi^{(2\nu+2)} + \cdots)\,. \tag{19.43}$$

The recursion relations then uniquely express all coefficients in terms of the free coefficients $\phi^{(0)}$ and $\phi^{(2\nu)}$. As an example, we note for later purposes that (for any value of ν)

$$\phi^{(2)} = \frac{1}{4(\nu-1)}\Box^{(0)}\phi^{(0)}\,, \tag{19.44}$$

where $\Box^{(0)}$ is the scalar wave operator defined by $\gamma^{(0)}$ on ∂M. Dimensional analysis shows that the higher coefficients $\phi^{(n)}$ for integer $n < 2\Delta_+ - d$ involve n derivatives of $\phi^{(0)}$.

We now couple our scalar to dynamical gravity using

$$S = S_{\text{grav}} + S_\phi^{\text{Bulk}}\,, \tag{19.45}$$

where S_{grav} is the action for gravity. We will postpone a discussion of boundary terms to Sect. 19.2; for now, we simply focus on solving the resulting equations of motion

$$R_{\mu\nu} - \tfrac{1}{2}Rg_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi G T^{(\text{matter})}_{\mu\nu}\,. \tag{19.46}$$

As in the vacuum case we write the metric in the form (19.17), and as in the solution for nondynamical gravity we write the scalar field as in (19.43). Note that we keep the logarithmic term in (19.21) for all d as, depending on the matter content, it may be necessary even for odd d. (When it is not needed, the equations of motion force its coefficient $\tilde{\gamma}_d$ to vanish.) The stress tensor of the scalar field then behaves like

$$\begin{aligned} &T^{(\text{matter})}_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu \\ &= \Delta_- z^{2(\Delta_- - 1)} \\ &\quad\times\left[\frac{\mathrm{d}}{2}(\phi^{(0)})^2\,\mathrm{d}z^2 + z\phi^{(0)}\partial_i\phi^{(0)}\,\mathrm{d}z\,\mathrm{d}x^i \right.\\ &\qquad \left. + \nu(\phi^{(0)})^2\gamma^{(0)}_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j + \cdots\right]. \end{aligned} \tag{19.47}$$

For $\Delta_- < 0$ and $\phi^{(0)} \neq 0$, the matter stress tensor turns out to diverge too rapidly at $z = 0$ for the equations of motion to admit an AlAdS solution. So for $\Delta_- < 0$ the only scalar field boundary condition consistent with the desired physics is $\phi^{(0)} = 0$. However, for $\Delta_- \geq 0$ the equations of motion *do* admit AlAdS solutions with $\phi^{(0)} \neq 0$ and further input is required to determine the boundary conditions. We will return to this issue in Sect. 19.2.2.

Evidently, the equations of motion admit solutions of the forms (19.17) and (19.43) only if the components of the matter stress tensor in Fefferman–Graham coordinates diverge as $1/z^2$ or slower. This result allows us to generalize our definition of asymptotically locally AdS spacetimes to include matter: an AlAdS spacetime with matter is a manifold M as above with fields satisfying the equations of motion and the requirement that $\Omega^2 T_{\mu\nu}$ admits a continuous limit to ∂M.

19.2 Variational Principles and Charges

Noether's theorem teaches us that variational principles provide a powerful link between symmetries and conservation laws, allowing the latter to be derived without detailed knowledge of the equations of motion. This procedure works as well for gravitational theories as for systems defined on a fixed spacetime background, though there is one additional subtlety. In more familiar theories, it is often sufficient to consider only variations of compact support so that all boundary terms arising from variations of an action can be discarded. However, as shown in Chap. 17 in the asymptotically flat context, when the gravitational constraints (which are just certain equations of motion!) are satisfied the gravitational charges become pure boundary terms with no contributions from the bulk. Discarding all boundary terms in Noether's theorem would thus lead to trivial charges and we will instead need to treat boundary terms with care. It is in part for this reason that we refer to *variational principles* as opposed to mere actions, the distinction being that all variations of the former vanish when the equations of motion and boundary conditions hold, even including any boundary terms that may arise in computing the variations. Constructing a good variational principle generally requires that we add boundary terms to the familiar bulk action, and that we tailor the choice of such boundary terms to the boundary conditions we wish to impose on ∂M.

19.2.1 A Toy Model of AdS: Gravity in a Box

We have seen that AlAdS spacetimes are conformally equivalent to manifolds with time-like boundaries. This means that (with appropriate boundary conditions) light signals can bounce off of ∂M and return to the interior in finite time, boundary conditions are needed for time evolution, and indeed much of physics in AlAdS spacetimes is indeed like field theory in a finite-sized box. This analogy also turns out to hold for the study of conservation laws in theories with dynamical gravity. It will, therefore, prove useful to first study conservation laws for gravity on a manifold M with a finite-distance time-like boundary ∂M, which will serve as a toy model for AlAdS gravitational dynamics. This subject, which we call *gravity in a box*, was historically studied for its own sake by *Brown* and *York* [19.23]. We largely follow their approach below. For simplicity we will assume that ∂M is globally hyperbolic with compact Cauchy surfaces as shown in Fig. 19.3, although the more general case can typically be treated by imposing appropriate boundary conditions in the asymptotic regions of ∂M.

Out first task is to construct a good variational principle. However, noted above this will generally require us to add boundary-condition-dependent boundary terms to the bulk action. It is thus useful to have some particular boundary condition (or, at least, a class of such conditions) in mind before we begin. In scalar field theory, familiar classes of boundary conditions include the Dirichlet condition ($\phi|_{\partial M}$ fixed, so $\delta\phi|_{\partial M} = 0$), the Neumann condition (which fixes the normal derivative), or the more general class of Robin conditions (which fix a linear combination of the two). All of these have analogs for our gravity in a box system, but for simplicity we will begin with a Dirichlet-type condition. Recall from Chap. 16 that, when discussing the initial value problem, the natural initial data on a Cauchy surface consists of the induced metric and the extrinsic curvature (or, equivalently, the conjugate momentum as described in Chap. 17). Since the equations of motion are covariant, the analysis of possible boundary conditions on time-like boundaries turns out to be very similar, so that the natural Dirichlet-type condition is to fix the induced metric h_{ij} on ∂M.

An important piece of our variational principle will, of course, be the Einstein–Hilbert action $S_{\mathrm{EH}} = \frac{1}{2\kappa}\int \sqrt{-g}\, R$ (with $\kappa = 8\pi G$). However, S_{EH} is not sufficient by itself as a standard calculation gives

$$\begin{aligned} \delta S_{\mathrm{EH}} &= \delta\left(\frac{1}{2\kappa}\int_M \sqrt{-g}R\right) \\ &= \frac{-1}{2\kappa}\int_M \sqrt{-g}\left(R^{\mu\nu} - \tfrac{1}{2}Rg^{\mu\nu}\right)\delta g_{\mu\nu} \\ &\quad + \frac{1}{2\kappa}\int_{\partial M} \sqrt{|h|}\hat{r}_\lambda G^{\mu\nu\rho\lambda}\nabla_\rho \delta g_{\mu\nu}\,, \end{aligned} \tag{19.48}$$

where $\hat{r}^\lambda$ is the outward pointing unit normal to ∂M and

$$G^{\mu\nu\rho\lambda} = g^{\mu(\rho}g^{\lambda)\nu} - g^{\mu\nu}g^{\rho\lambda}\,. \tag{19.49}$$

In (19.48) we have discarded boundary terms not associated with ∂M (i. e., boundary terms in any asymptotic regions of M) as they will play no role in our analysis. Nevertheless, the second term in (19.48) (the boundary term) generally fails to vanish for useful boundary conditions, so that S_{EH} is not fully stationary on solutions.

However, when $\delta h_{ij} = 0$ this problem term turns out to be an exact variation of another boundary term,

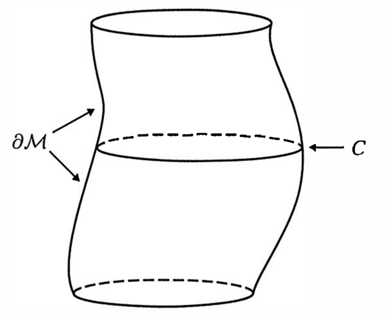


Fig. 19.3 A sketch of the spacetime $\mathcal{M}$. The codimension two surface C is a Cauchy surface of the boundary ∂M

known as the Gibbons–Hawking term, given by the integral of the trace of the extrinsic curvature of ∂M. (For related reasons the addition of this term is necessary when constructing a gravitational path integral, see [19.24]). As a result, enforcing the boundary condition $\delta h_{ij} = 0$ guarantees that all variations of the action

$$S_{\text{Dirichlet in a box}} = S_{\text{EH}} + S_{\text{GH}} = \frac{1}{2\kappa}\int_{\mathcal{M}} \sqrt{-g}R - \frac{1}{\kappa}\int_{\partial M} \sqrt{|h|}K\,, \tag{19.50}$$

where $K = h_{ij}K^{ij}$ is the trace of the extrinsic curvature on ∂M, vanish precisely when the bulk equations of motion hold. Thus (19.50) gives a good variational principle for our Dirichlet problem.

Now, Noether's theorem teaches us that every continuous symmetry of our system should lead to a conservation law (although the conservation laws associated with pure gauge transformations are trivial). Gravity in a box is defined by the action (19.50) and by the choice of some Lorentz-signature metric h_{ij} on ∂M. The first ingredient, the action (19.50), is manifestly invariant under any diffeomorphisms of M. Such diffeomorphisms are generated by vector fields ξ on M that are tangent to ∂M at the boundary (so that the diffeomorphism maps ∂M to itself). As before, we use $\hat{\xi}$ to denote the induced vector field on ∂M. The associated diffeomorphism of M will preserve h_{ij} if $\hat{\xi}$ is a Killing field on the boundary. As discussed in Chap. 17, a diffeomorphism supported away from the boundary should be pure gauge. So it is natural to expect that the asymptotic symmetries of our system are classified by the choice of boundary Killing field $\hat{\xi}$, with the particular choice of a bulk extension ξ being pure gauge.

This set-up should remind the reader of (nongravitational) field theories on fixed spacetime backgrounds. There one finds conservation laws associated with each Killing field of the background metric. Here again the conservation laws are associated with Killing fields of the background structure, although now the only such structure is the boundary metric h_{ij}.

Pursuing this analogy, let us recall the situation for field theory on a fixed (nondynamical) spacetime background. There, Noether's theorem for global symmetries (e.g., translations along some Killing field ξ_{KVF}) would instruct us to vary the action under a space-time generalization of the symmetry (e.g., diffeomorphism along $f(x)\xi_{\text{KVF}}$ for general smooth functions $f(x)$, or more generally under arbitrary diffeomorphisms). It is clear that the analog for gravity in a box is just to vary (19.50) under a general diffeomorphism of M.

It turns out to be useful to do so in two steps. Let us first compute an arbitrary variation of (19.50). By construction, it must reduce to a boundary term when the equations of motion hold, and it must vanish when $\delta h_{ij} = 0$. Thus it must be linear in δh_{ij}. A direct calculation [19.12, Appendix E]) gives

$$\delta S_{\text{Dirichlet in a box}} = \frac{1}{2}\int_{\partial M} \sqrt{|h|}\tau^{ij}\delta h_{ij}\,, \tag{19.51}$$

where $\tau^{ij} = \kappa^{-1}(K^{ij} - Kh^{ij})$. This τ^{ij} is sometimes referred to as the radial conjugate momentum since it has the same form as the (undensitized) conjugate momentum introduced on space-like surfaces in Chap. 17. This agreement, of course, follows from general principles of Hamilton–Jacobi theory. The reader should recall that for field theory in a fixed spacetime background the functional derivative of the action with respect to the metric defines the field theory stress tensor. By analogy, the object τ^{ij} defined above is often called the boundary stress tensor (or the Brown–York stress tensor) of the gravitational theory.

Let us now specialize to the case where our variation is a diffeomorphism of M. As we have seen, ξ also induces a diffeomorphism of the boundary ∂M generated by some $\hat{\xi}$. Then $\delta h_{ij} = D_i\hat{\xi}_j + D_j\hat{\xi}_i$, where D_i is the covariant derivative compatible with h_{ij}. Using the symmetry of $\tau^{ij} = \tau^{ji}$ we find

$$\delta S_{\text{Dirichlet in a box}} = \int_{\partial M} \sqrt{|h|}\tau^{ij}D_i\hat{\xi}_j = -\int_{\partial M} \sqrt{|h|}\hat{\xi}_j D_i\tau^{ij}\,, \tag{19.52}$$

where in the last step we integrate by parts and take $\hat{\xi}$ to have compact support on ∂M so that we may discard any boundary terms. Since $\hat{\xi}$ is otherwise arbitrary, we conclude that

$$D_i \tau^{ij} = 0 ; \tag{19.53}$$

i. e., τ^{ij} is covariantly conserved on ∂M when the equations of motion hold in the bulk. In fact, since τ^{ij} is the radial conjugate momentum, it should be clear from Chap. 17 that (19.53) can also be derived directly from the equations of motion by evaluating the radial-version of the diffeomorphism constraint on ∂M. (The radial version of the Hamiltonian constraint imposes another condition on τ^{ij} that can be used to determine the trace $\tau = \tau^{ij} h_{ij}$ in terms of the traceless part of τ^{ij}.)

If we now take $\hat{\xi}$ to be a boundary Killing field, we find $D_i(\tau^{ij}\hat{\xi}_j) = 0$, so that the so-called Brown–York charge

$$Q_{\text{BY}}[\xi] := -\int_C \sqrt{q} n_i \tau^{ij} \hat{\xi}_j \tag{19.54}$$

is independent of the choice of Cauchy surface C in ∂M. Here n_i is a unit future-pointing normal to C and $\sqrt{q}$ is the volume element induced on C by h_{ij}. Although these charges were defined by methods quite different from the Hamiltonian techniques of Chap. 17, we will argue in Sect. 19.3 below that the end result is identical up to a possible choice of zero-point. Once again, the argument will turn out to be essentially the same as one would give for field theory in a fixed nondynamical background.

Before proceeding to the AdS case, let us take a moment to consider other possible boundary conditions. We see from (19.51) that the action (19.50) also defines a valid variational principle for the boundary condition $\tau^{ij} = 0$. Of course, with this choice the charges (19.54) all vanish. But this should be no surprise. Since the condition $\tau^{ij} = 0$ is invariant under *all* diffeomorphisms of M, there is no preferred subset of nontrivial asymptotic symmetries; all diffeomorphisms turn out to generate pure gauge transformations. One may also study more complicated boundary conditions by adding additional boundary terms to the action (19.50), although we will not pursue the details here.

19.2.2 Variational Principles for Scalar Fields in AdS

As the reader might guess, our discussion of AlAdS gravity will follow in direct analogy to the above treatment of gravity in a box. Indeed, the only real difference is that we must work a bit harder to construct a good variational principle. We will first illustrate the relevant techniques below by constructing a variational principle for a scalar field on a fixed AdS background, after which we will apply essentially identical techniques to AdS gravity itself in Sect. 19.2.3.

We will construct our variational principle using the so-called counterterm subtraction approach pioneered in [19.25, 26] and further developed in [19.17, 18]. Our discussion below largely follows [19.17], with minor additions from [19.11]. We begin with the bulk action S_ϕ^{Bulk} of (19.40) and compute

$$\delta S_\phi^{\text{Bulk}} = -\int_{\partial M} \sqrt{|h|} \hat{r}^\mu \partial_\mu \phi \delta\phi , \tag{19.55}$$

where $\hat{r}^\mu$ is the outward-pointing unit normal to ∂M so that $\hat{r}^\mu \partial_\mu = -\frac{z}{\ell}\partial_z$. The form of (19.55) might appear to suggest that S_ϕ^{Bulk} defines a good variational principle for any boundary condition that fixes ϕ on ∂M. However, the appearance of inverse powers of z means that we must be more careful, and that S_ϕ^{Bulk} will suffice only when $\delta\phi$ vanishes sufficiently rapidly.

It is, therefore, useful to write (19.55) in terms of the finite coefficients $\phi^{(2n)}, \phi^{(2(\nu+n))}$ of (19.42) (or the corresponding coefficients in (19.43)). The exact expression is not particularly enlightening, and for large ν there are many singular terms to keep track of. What is useful to note, however, is that all of the singular terms turn out to be exact variations. In particular, using (19.44) one may show for noninteger $\nu < 2$ that the action

$$S_\phi = S_\phi^{\text{Bulk}} + \int_{\partial M} \sqrt{|h|} \times \left(-\frac{\Delta_-}{2\ell}\phi^2 + \frac{\ell}{4(\nu-1)} h^{ij}\partial_i\phi\partial_j\phi \right) \tag{19.56}$$

satisfies

$$\delta S_\phi = 2\nu \ell^{d-1} \int_{\partial M} \sqrt{|\gamma^{(0)}|} \phi^{(2\nu)} \delta\phi^{(0)} . \tag{19.57}$$

Since the boundary terms in (19.56) are each divergent in and of themselves, they are known as counterterms in analogy with the counterterms used to cancel ultraviolet divergences in quantum field theory. These divergences cancel against divergences in S_Φ^{Bulk} and the full action S_Φ is finite for any field of the form (19.42) with noninteger $\nu < 2$. Similar results hold for noninteger $\nu > 2$ if additional higher-derivative boundary terms are included in (19.56). We will comment on differences for integer ν at the end of this section.

It is clear that S_Φ provides a good variational principle so long as the boundary conditions either fix $\phi^{(0)}$ or set $\phi^{(2\nu)} = 0$. We may now identify

$$\Phi_{\mathrm{bndy}} := 2\nu\ell^{d-1}\phi^{(2\nu)} \tag{19.58}$$

as an AdS scalar response function analogous to the boundary stress tensor τ^{ij} introduced in Sect. 19.2.1. Note that adding an extra boundary term $\int \sqrt{\gamma^{(0)}}W[\phi^{(0)}]$ to S_Φ allows one to instead use the Robin-like boundary condition

$$\phi^{(2\nu)} = -\frac{\ell}{2\nu}W'[\phi^{(0)}]\,, \tag{19.59}$$

where W' denotes the derivative of W with respect to its argument.

Recall from Sect. 19.1.6 that requiring the energy to be bounded below restricts ν to be real (in which case we take ν nonnegative). That there are further implications for large ν can also be seen from (19.56). Note that the final term in (19.56) is a kinetic term on ∂M and that for $\nu > 1$ it has a sign *opposite* to that of the bulk kinetic term. Counting powers of z shows that this boundary kinetic term vanishes at ∂M for $\nu < 1$, but contributes for $\nu > 1$. In this case, for any perturbation that excites $\phi^{(0)}$ and which is supported sufficiently close to ∂M, the boundary kinetic term in (19.56) turns out to be more important than the bulk kinetic term. Thus the perturbation has negative kinetic energy. One says that the theory contains ghosts, and any conserved energy is expected to be unbounded below [19.11]. For this reason, for $\nu > 1$ one typically allows only boundary conditions that fix $\phi^{(0)}$. Of course, as noted in Sect. 19.2.2, for $\nu > d/2$ coupling the theory to dynamical gravity and requiring the spacetime to be AlAdS will further require $\phi^{(0)} = 0$. On the other hand, for real $0 < \nu < 1$ all of the above boundary conditions lead to ghost-free scalar theories.

The story of noninteger $\nu > 2$ is much the same as that of $\nu \in (1,2)$. Adding additional higher-derivative boundary terms to (19.56) again leads to an action that satisfies (19.57). While one can find actions compatible with general boundary conditions (19.59), the only ghost-free theories fix $\phi^{(0)}$ on ∂M. The story of integer ν is more subtle; the factors of $\ln z$ arising in that case from (19.43) mean that we can find a good variational principle only by including boundary terms that depend explicitly on the defining function Ω of the chosen conformal frame. Doing so again leads to ghosts unless $\phi^{(0)}$ is fixed as a boundary condition [19.11].

19.2.3 A Variational Principle for AlAdS Gravity

We are now ready to construct our variational principle for AlAdS gravity. As for the scalar field above, we will start with a familiar bulk action and then add boundary terms. One may note that in the scalar case our final action (19.56) consists essentially of adding boundary terms to S_Φ^{Bulk} which i) are written as integrals of local scalars built from ϕ and its tangential derivatives along ∂M and ii) precisely cancel divergent terms in S_Φ^{Bulk}. This motivates us to follow the strategy of [19.18] for the gravitational case in which we first identify divergent terms in a familiar action and write these terms as local scalars on ∂M. We may then construct a finite so-called renormalized action by adding boundary counterterms on ∂M to cancel the above divergences. At the end of this process we may check that this renormalized action yields a good variational principle for interesting boundary conditions. In analogy with Sect. 19.2.1, for simplicity in the remainder of this chapter we take the induced (conformal) metric on ∂M to be globally hyperbolic with compact Cauchy surfaces.

Let us begin with an action containing the standard Einstein–Hilbert and cosmological constant terms in the bulk, along with the Gibbons–Hawking term. It will facilitate our discussion of divergent terms to consider a regulated action in which the boundary has effectively been moved in to $z = \epsilon$. For the moment, we choose some $\epsilon_0 > \epsilon$ and impose the Fefferman–Graham gauge (19.17) for all $z < \epsilon_0$, so that this gauge holds in particular at the regulated boundary. This gauge fixing at finite z is merely an intermediate step to simplify the analysis. We will be able to loosen this condition once we have constructed the final action. We let $h_{ij} = (\ell/z)^2\gamma_{ij}|_{z=\epsilon}$ be the induced metric on this regu-

lated boundary and study the action

$$S_{\text{reg}} = \frac{1}{2\kappa}\int_{z\geq\epsilon} \sqrt{|g|}(R-2\Lambda) - \frac{1}{\kappa}\int_{z=\epsilon}\sqrt{|h|}K$$
$$= \frac{-\ell^{d-1}}{2\kappa}\int_{z=\epsilon}\sqrt{|\gamma^{(0)}|}$$
$$\times\Big(\epsilon^{-d}a_{(0)} + \epsilon^{-d+2}a_{(2)} + \ldots$$
$$+\epsilon^{-2}a_{(d-2)} - \log(\epsilon^2)a_{(d)}\Big)$$
$$+ (\text{finite})\,, \tag{19.60}$$

where $K = h_{ij}K^{ij}$ is the trace of the extrinsic curvature of the regulated boundary ∂M_ϵ at $z=\epsilon$ and the form of the divergences follows from (19.21). The coefficient $a_{(d)}$ vanishes for odd d. For even d it is called the conformal anomaly for reasons to be explained below.

In analogy with the scalar field results of Sect. 19.2.2, one finds that the coefficients $a_{(n)}$ which characterize the divergent terms are all local scalars built from $\gamma^{(0)}_{ij}$ and its derivatives along ∂M. This follows directly from the fact that all terms $\gamma^{(n)}$ with $n \leq d$ in the Fefferman–Graham expansion (19.21) are local functions of $\gamma^{(0)}_{ij}$ and its derivatives along ∂M. Dimensional analysis shows that $a_{(n)}$ involves precisely $2n$ derivatives and the detailed coefficients $a_{(n)}$ can be found to any desired order by direct calculation. For example, for $n \neq d$ the $a_{(n)}$ are given by (see e.g. [19.18])

$$a_{(0)} = -2(d-1)\,,\quad a_{(2)} = \frac{(d-4)\mathcal{R}}{2(d-2)}\,,$$
$$a_{(4)} = -\frac{d^2-9d+16}{4(d-4)}$$
$$\times\left(\frac{d\mathcal{R}^2}{4(d-2)^2(d-1)} - \frac{\mathcal{R}^{ij}\mathcal{R}_{ij}}{(d-2)^2}\right),\quad \ldots\,, \tag{19.61}$$

where as in Sect. 19.1.4, $\mathcal{R}$ and $\mathcal{R}_{ij}$ are the Ricci scalar and Ricci tensor of $\gamma^{(0)}$ on ∂M. For $d = 2$, 4, the log terms are given by

$$d=2:\quad a_{(2)} = \frac{-\mathcal{R}}{2}\,,$$
$$d=4:\quad a_{(4)} = \left(\frac{\mathcal{R}^2}{24} - \frac{\mathcal{R}^{ij}\mathcal{R}_{ij}}{8}\right). \tag{19.62}$$

As foreshadowed above, we now define the renormalized action

$$S_{\text{ren}} = \lim_{\epsilon\to 0}(S_{\text{reg}} + S_{\text{ct}})\,, \tag{19.63}$$

where

$$S_{\text{ct}} := \frac{\ell^{d-1}}{2\kappa}\int_{z=\epsilon}\sqrt{-\gamma^{(0)}}$$
$$\times\Big(\epsilon^{-d}a_{(0)} + \epsilon^{-d+2}a_{(2)} + \ldots$$
$$+\epsilon^{-2}a_{(d-2)} - \log(\epsilon^2)a_{(d)}\Big) \tag{19.64}$$

is constructed to precisely cancel the divergent terms in S_{ren}. The representation (19.64) makes the degree of divergence in each term manifest. However, the use of ϵ in defining S_{ct} suggests a stronger dependence on the choice of defining function Ω (and thus, on the choice of conformal frame) than is actually the case. To understand the true dependence, we should use the Fefferman–Graham expansion to instead express S_{ct} directly in terms of the (divergent) metric h induced on ∂M by the unrescaled bulk metric g as was done in [19.26]. Dimensional analysis and the fact that each $a_{(n)}$ involves precisely $2n$ derivatives shows that this removes all explicit dependence on ϵ save for the logarithmic term in even d. In particular, formally taking ϵ to zero we may write

$$S_{\text{ct}} = \frac{\ell}{2\kappa}\int_{\partial M}\sqrt{|h|}$$
$$\times\left[-\frac{2(d-1)}{\ell^2} - \frac{\mathcal{R}_h}{(d-2)} + \ldots\right.$$
$$\left.- \frac{\epsilon^d\log(\epsilon^2)a_{(d)}}{\ell^2}\right], \tag{19.65}$$

where the $\mathcal{R}_h$ (Ricci scalar of h) term only appears for $d \geq 3$ and the dots represent additional terms that appear only for $d \geq 5$.

In general, the coefficients in (19.65) differ from those in (19.60) due to subleading divergences in a given term in (19.65) contributing to the coefficients of seemingly lower-order terms in (19.60). However, the logarithmic term has precisely the same coefficient $a_{(d)}$ in both (19.65) and (19.60). Since the logarithmic term in (19.21) is multiplied by z^d, only the leading $-\frac{2(d-1)}{\ell^2}\sqrt{|h|}$ term in (19.65) could contribute to

any discrepancy. However, the first variation of a determinant is a trace, and the trace of the logarithmic coefficient $\bar{\gamma}^{(d)}_{ij}$ vanishes by (19.22).

Thus for d odd (where the log term vanishes) the renormalized action $S_{\rm ren}$ can be expressed in a fully covariant form in terms of the physical metric g; all dependence on the defining function Ω (and so on the choice of conformal frame) has disappeared. We, therefore, now drop the requirement that any Fefferman–Graham gauge be imposed for odd d. However, for even d, the appearance of $\log(\epsilon^2)$ in (19.65) indicates that $S_{\rm ren}$ does, in fact, depend on the choice of defining function Ω (and thus on the choice of conformal frame). In analogy with quantum field theory, this dependence is known as the conformal anomaly. By replacing ϵ with Ω in (19.65), we could again completely drop the requirement of Fefferman–Graham gauge in favor of making explicit the above dependence on Ω. However, an equivalent procedure is to require that the expansion (19.21) hold up through order $\gamma^{(d)}$ and to replace ϵ in (19.65) by the Fefferman–Graham coordinate z. We will follow this latter approach (which is equivalent to imposing Fefferman–Graham gauge only on the stated terms in the asymptotic expansion) as it is more common in the literature.

We are finally ready to explore variations of $S_{\rm ren}$. Since $S_{\rm ren}$ was constructed by adding only boundary terms to the usual bulk action, we know that $\delta S_{\rm ren}$ must be a pure boundary term on solutions. As before, we will discard boundary terms in the far past and future of M and retain only the boundary term at ∂M. Since ∂M is globally hyperbolic with compact Cauchy surfaces, performing integrations by parts on ∂M will yield boundary terms only in the far past and future of ∂M. Discarding these as well allows us to write

$$\delta S_{\rm ren} = \int_{\partial M} S^{\mu\nu}\delta g_{\mu\nu}\,, \qquad (19.66)$$

for some $S^{\mu\nu}$. However, let us now return to the Fefferman–Graham gauge and use it to expand $\delta g_{\mu\nu}$, as in (19.21). Since $S_{\rm ren}$ is finite, $\delta S_{\rm ren}$ must be finite as well, but the leading term in $\delta g_{\mu\nu}$ is of order z^{-2}. So the leading term in $S_{\mu\nu}$ must be of order z^2. It follows that only these leading terms can contribute to (19.66). Since the leading term in $\delta g_{\mu\nu}$ involves $\delta\gamma^{(0)}_{ij}$, we may write

$$\delta S_{\rm ren} = \tfrac{1}{2}\int_{\partial M} \sqrt{|\gamma^0|}\, T^{ij}_{\rm bndy}\delta\gamma^{(0)}_{ij} \qquad (19.67)$$

for some finite so-called boundary stress tensor $T^{ij}_{\rm bndy}$ on ∂M. For odd d, the fact that $S_{\rm ren}$ is invariant under arbitrary changes of conformal frame $\delta\gamma^{(0)}_{ij} = e^{-2\sigma}\gamma^{(0)}_{ij}$ immediately implies that the boundary stress tensor is traceless: $T_{\rm bndy} := \gamma^{(0)}_{ij}T^{ij}_{\rm bndy} = 0$. In even dimensions, the trace is determined by the conformal anomaly of $S_{\rm ren}$ (i. e., by the logarithmic term in either (19.60) or (19.65)) and one finds $T_{\rm bndy} = -\ell^{d-1}a_{(d)}/\kappa$. This result may also be derived by considering the radial version of the Hamiltonian constraint from Chap. 17 and evaluating this constraint at ∂M.

Comparing with Sect. 19.2.1, it is clear that we may write

$$T^{ij}_{\rm bndy} = \lim_{\epsilon\to 0}\left(\frac{\ell}{\epsilon}\right)^{d+2}\left(\tau^{ij} + \tau^{ij}_{ct}\right), \qquad (19.68)$$

where again $\tau_{ij} = \kappa^{-1}(K_{ij} - Kh_{ij})$ and the new term $\tau^{ij}_{\rm ct}$ comes from varying $S_{\rm ct}$. In Fefferman–Graham gauge one finds by explicit calculation that for d odd

$$T^{ij}_{\rm bndy} = \frac{d\ell^{d-1}}{2\kappa}\gamma^{(d)\,ij}\,. \qquad (19.69)$$

For d even there are extra contributions associated with the conformal anomaly, which are thus all determined by $\gamma^{(0)}$; e.g. [19.18]

- For $d = 2$:

$$T^{ij}_{\rm bndy} = \frac{\ell}{\kappa}\left(\gamma^{(2)ij} + \tfrac{1}{2}\mathcal{R}\gamma^{(0)ij}\right), \qquad (19.70)$$

- For $d = 4$:

$$T^{ij}_{\rm bndy} = \frac{2\ell^3}{\kappa}\Big[\gamma^{(4)ij} - \tfrac{1}{8}\left((\gamma^{(2)})^2 - \gamma^{(2)kl}\gamma^{(2)}_{kl}\right)\gamma^{(0)ij} - \tfrac{1}{2}\gamma^{(2)ik}\gamma^{(2)\,j}_{\ \ k} + \tfrac{1}{4}\gamma^{(2)}\gamma^{(2)ij} + \frac{3}{2}\bar{\gamma}^{(4)ij}\Big], \qquad (19.71)$$

where $\gamma^{(2)}$, $\bar{\gamma}^{(4)}$ are given by (19.19), (19.23), (19.24). In all cases, we see that we may use $\gamma^{(0)}_{ij}$, $T^{ij}_{\rm bndy}$ to parametrize the free data in the Fefferman–Graham expansion.

The reader should note that the particular value of $T^{ij}_{\rm bndy}$ on a given solution depends on the choice of a representative $\gamma^{(0)}$ and thus on the choice of conformal frame. For d odd this dependence is a simple scaling, although it is more complicated for d even.

Yet this does not diminish the utility of T^{ij}_{bndy}. For example, we see immediately from (19.67) that S_{ren} defines a good variational principle whenever i) $\gamma^{(0)}$ is fixed as a boundary condition or ii) d is odd, so that T^{ij}_{bndy} is traceless, and we fix only the conformal class of $\gamma^{(0)}$.

We close this section with some brief comments on other possible boundary conditions. We can see from (19.67) that S_{ren} is also a good variational principle if we fix $T^{ij}_{\text{bndy}} = 0$. As in Sect. 19.2.2, one may obtain variational principles for more complicated boundary conditions by adding further finite boundary terms to (19.65); see [19.27] for details. However, just as for scalar fields with $\nu > 1$, boundary conditions that allow $\gamma^{(0)}$ to vary generally lead to ghosts [19.11] (with the exception that for d odd no ghosts arise from allowing $\gamma^{(0)}$ to vary by a conformal factor). For this reason we consider below only boundary conditions that fix $\gamma^{(0)}$, or at least its conformal class for d odd.

19.2.4 Conserved Charges for AlAdS Gravity

We are now ready to apply the Brown–York-type procedure discussed in Sect. 19.2.1 to construct conserved charges for AlAdS gravity. The key step is again an argument analogous to (19.52) to show conservation of T^{ij}_{bndy} on ∂M. We give the derivation here in full to highlight various subtleties of the AdS case. We also generalize the result slightly by coupling the AlAdS gravity theory of Sect. 19.2.3 to the scalar theory of Sect. 19.2.2. For definiteness we assume that the boundary conditions fix both $\gamma^{(0)}$ and $\phi^{(0)}$ (up to conformal transformations $(\gamma^{(0)}_{ij}, \phi^{(0)}) \to (e^{-2\sigma}\gamma^{(0)}_{ij}, e^{\Delta-\sigma}\phi^{(0)})$) for odd d, where the transformation of $\phi^{(0)}$ is dictated by (19.42) and we take ν noninteger so that no log terms arise from the scalar field. However, the more general case is quite similar [19.10, 27].

We thus consider the action $S_{\text{total}} = S_{\text{ren}} + S_\phi$. The reader should be aware that, because the counterterms in S_ϕ explicitly depend on the boundary metric $\gamma^{(0)}$, this coupling to matter will change certain formulae in Sect. 19.2.3. In particular, if we now make the natural definition

$$T^{ij}_{\text{bndy}} = \frac{2}{\sqrt{|\gamma^{(0)}|}} \frac{\delta S_{\text{total}}}{\delta \gamma^{(0)}_{ij}}, \tag{19.72}$$

varying the action under a boundary conformal transformation leads to the more general condition

$$T_{\text{bndy}} - \Delta_- \Phi_{\text{bndy}} \phi^{(0)} = -\frac{\ell^{d-1} a_{(d)}}{\kappa}, \tag{19.73}$$

which reduces to the trace constraint of Sect. 19.2.3 only for $\Phi_{\text{bndy}} = 0$, $\phi^{(0)} = 0$, or $\Delta_- = 0$. Recall that Φ_{bndy} is given by (19.58).

The coupling to S_ϕ similarly modifies the divergence condition (19.52) of Sect. 19.2.1. Using the definition (19.72), we find

$$\delta S_{\text{total}} = \int_{\partial M} \sqrt{|\gamma^{(0)}|} \left(\tfrac{1}{2} T^{ij}_{\text{bndy}} \delta \gamma^{(0)}_{ij} + \Phi_{\text{bndy}} \delta \phi^{(0)} \right). \tag{19.74}$$

Let us consider the particular variation associated with a bulk diffeomorphism ξ. It is sufficient here to consider bulk diffeomorphisms compatible with whatever defining function Ω we have used to write (19.74); i.e., for which $\pounds_\xi \Omega = 0$. As described in Sect. 19.1.5, other diffeomorphisms differ only in that they also induce a change of conformal frame. Since we already extracted the information about T^{ij}_{bndy} (and in particular, about its trace) that can be obtained by changing conformal frame in Sect. 19.2.3; we lose nothing by restricting here to vector fields with $\pounds_\xi \Omega = 0$.

As described in Sect. 19.1.5, we then find $\delta\gamma^{(0)} = \pounds_{\hat\xi} \gamma^{(0)}$, $\delta\phi^{(0)} = \pounds_{\hat\xi} \phi^{(0)}$, where $\hat\xi$ is the vector field induced by ξ on ∂M. Thus (19.74) reads

$$\begin{aligned} \delta_\xi S_{\text{ren}} = 0 &= \int_{\partial M} \sqrt{|\gamma^{(0)}|} \left(T^{ij} D_i \hat\xi_j + \frac{\delta S_{\text{ren}}}{\delta \phi^{(0)}} \pounds_{\hat\xi} \phi^{(0)} \right) \\ &= -\int_{\partial M} \sqrt{|\gamma^{(0)}|}\, \hat\xi_j \left(D_i T^{ij} - \Phi_{\text{bndy}} D^j \phi^{(0)} \right), \end{aligned} \tag{19.75}$$

where D_i is again the covariant derivative on ∂M compatible with $\gamma^{(0)}$, all indices are raised and lowered with $\gamma^{(0)}$, and we have dropped the usual surface terms in the far past and future of ∂M. Recalling that all $\hat\xi^i$ can arise from bulk vector fields ξ compatible with any given Ω, we see that (19.75) must hold for any $\hat\xi_j$. Thus,

$$D_i T^{ij}_{\text{bndy}} = \Phi_{\text{bndy}} D^j \phi^{(0)}; \tag{19.76}$$

i.e., T^{ij}_{bndy} is conserved on ∂M up to terms that may be interpreted as scalar sources. These sources are analogous to sources for the stress tensor of, say, a scalar field on a fixed spacetime background when the scalar field is also coupled to some background potential. Here the role of the background potential is played by $\phi^{(0)}$, which we have fixed as a boundary condition. As in Sect. 19.2.1, the divergence condition (19.76) may also

be derived from the radial version of the diffeomorphism constraint from Chap. 17 evaluated on ∂M. For $\phi^{(0)} = 0$ and d odd one immediately arrives at (19.20) using (19.76) and (19.69).

We wish to use (19.76) to derive conservation laws for asymptotic symmetries. Here it is natural to say that a diffeomorphism ξ of M is an asymptotic symmetry if the there is *some* conformal frame in which the induced vector field $\hat{\xi}$ on ∂M is i) a Killing field of $\gamma^{(0)}$ and ii) a solution of $\pounds_{\hat{\xi}}\phi^{(0)} = 0$. Due to the transformations of $\gamma^{(0)}, \phi^{(0)}$ under boundary conformal transformations, this is completely equivalent to first choosing an arbitrary conformal frame and then requiring

$$\pounds_{\hat{\xi}}\gamma^{(0)}_{ij} = -2\sigma\gamma^{(0)}_{ij}\ , \quad \pounds_{\hat{\xi}}\phi^{(0)} = \Delta_{-}\sigma\phi^{(0)}\ . \tag{19.77}$$

The first requirement says that $\hat{\xi}$ is a conformal Killing field of $\gamma^{(0)}_{ij}$ with $\frac{1}{d}D_i\hat{\xi}^i = -\sigma$ and the second says that it acts on $\phi^{(0)}$ like the corresponding infinitesimal conformal transformation.

For even d, we must also preserve the boundary condition that $\gamma^{(0)}$ be fixed (even including the conformal factor) and the requirement of Sect. 19.2.3 that the Fefferman–Graham gauge hold to the first few orders in the asymptotic expansion. An analysis similar to that of Sect. 19.1.5 then shows that we must have $\xi^z = \frac{z}{d}D_i\hat{\xi}^i$ to leading order near ∂M. In particular, for $D_i\hat{\xi}^i \neq 0$ an asymptotic symmetry ξ must be noncompatible with Ω is just the right way to leave $\gamma^{(0)}$ invariant.

As a side comment, we mention that the trivial asymptotic symmetries (the pure gauge transformations) are just those with $\hat{\xi} = 0$. This means that they act trivially on both T^{ij}_{bndy} and Φ_{bndy} of Sect. 19.2.2, so that both T^{ij}_{bndy} and Φ_{bndy} are gauge invariant. This conclusion is obvious in retrospect, as these response functions are functional derivatives of the action with respect to the boundary conditions $\gamma^{(0)}_{ij}$ and $\phi^{(0)}$. Since both the action and any boundary conditions are gauge invariant by definition, so too must be the functional derivatives T^{ij}_{bndy} and Φ_{bndy}.

Returning to our construction of charges, note that for any asymptotic symmetries as above we may compute

$$\begin{aligned} D_i\left(T^{ij}_{\text{bndy}}\hat{\xi}_j\right) &= -\sigma\left(T_{\text{bndy}} - \Delta_{-}\Phi_{\text{bndy}}\phi^{(0)}\right) \\ &= \sigma\frac{\ell^{d-1}a_{(d)}}{\kappa}\ , \end{aligned} \tag{19.78}$$

where in the final step we have used (19.73).

In analogy with Sect. 19.2.1, we now consider the charges

$$Q[\xi] = -\int_C \sqrt{q}\, n_i T^{ij}_{\text{bndy}}\hat{\xi}_j\ , \tag{19.79}$$

where C is a Cauchy surface of ∂M, $\sqrt{q}$ is the volume element induced on C by $\gamma^{(0)}$, and n^i is the unit future pointing normal to C with respect to $\gamma^{(0)}$. It follows from (19.78) that these charges can depend on C only through a term built from the conformal anomaly $a_{(d)}$.

It is now straightforward to construct a modified charge $\tilde{Q}[\xi]$ which is completely independent of C. The essential point here is to recall that $a_{(d)}$ depends only on the boundary metric $\gamma^{(0)}$. Since we have fixed $\gamma^{(0)}$ as a boundary condition, the dependence on C is the same for any two allowed solutions. Thus on a given solution s we need only define

$$\tilde{Q}[\xi](s) = Q[\xi](s) - Q[\xi](s_0)\ , \tag{19.80}$$

where s_0 is an arbitrary reference solution satisfying the same boundary condition and which we use to set the zero-point. The construction (19.80) is sufficiently trivial so that one often refers to $Q[\xi]$ itself as being conserved.

Our construction of the charges $Q[\xi]$, $\tilde{Q}[\xi]$ depended on the choice of some conformal frame. However, it is easy to see that the charges are, in fact, independent of this choice for d odd. In that case, the factors $\sqrt{q}$, n_i, and T^{ij}_{bndy} all simply scale under a boundary conformal transformation, and dimensional analysis shows that the combination (19.79) is invariant. For even d there are additional terms in the transformation of T^{ij}_{bndy}. However, as usual these depend only on $\gamma^{(0)}$ so that they cancel between the two terms in (19.80). Thus even in this case for fixed s_0 the charges (19.80) are independent of the conformal frame.

To make the above procedure seem more concrete, we now quickly state results for the AdS_3 and AdS_4 Schwarzschild solutions

$$\begin{aligned} ds^2 = &-\left(1 - \frac{2c_d GM}{\rho^{d-2}} + \frac{\rho^2}{\ell^2}\right) d\tau^2 \\ &+ \frac{d\rho^2}{1 - \frac{2c_d GM}{\rho^{d-2}} + \frac{\rho^2}{\ell^2}} + \rho^2\, d\Omega^2_{(d-2)}\ , \end{aligned} \tag{19.81}$$

where $c_3 = 1$ and $c_4 = \frac{4}{3\pi}$. The boundary stress tensor may be calculated by converting to Fefferman–Graham

coordinates, say for the conformal frame defined by $\Omega = \rho^{-1}$. (Note that the Fefferman–Graham radial coordinate z will agree with ρ only at leading order.) One then finds the energy

$$Q[-\partial_\tau] = \begin{cases} M, & d = 3 \\ M + \dfrac{3\pi\ell^2}{32G}, & d = 4, \end{cases} \quad (19.82)$$

where we remind the reader that energies $E = -Q[\partial_\tau] = Q[-\partial_\tau]$ are conventionally defined in this way with an extra minus sign to make them positive. We see that for $d = 3$ we recover the expected result for the energy of the spacetime. For $d = 4$ we also recover the expected energy up to a perhaps unfamiliar choice of zero-point which we will discuss further in Sect. 19.3.4.

19.2.5 Positivity of the Energy in AlAdS Gravity

Thus far we have treated all charges $Q[\xi]$ on an equal footing. However, when $\hat{\xi}$ is everywhere time like and future-directed on ∂M, it is natural to call $E = Q[-\xi]$ an *energy* and to wonder whether E is bounded below. Such a result was established in Chap. 18 for the ADM energy of asymptotically flat spacetimes, and the Witten spinor methods [19.28, 29] discussed there generalize readily to asymptotically AdS (AAdS) spacetimes as long as the matter fields satisfy the dominant energy condition and decay sufficiently quickly at ∂M [19.30]. In particular, this decay condition is satisfied for the scalar field of Sect. 19.2.2 with $m^2 \geq m^2_{\rm BF}$ when $\phi^{(0)}$ is fixed as a boundary condition. Extensions to more general scalar boundary conditions can be found in [19.31–35]. Here the details of the boundary conditions are important, as boundary conditions for which the W of (19.59) diverges sufficiently strongly in the negative direction tend to make any energy unbounded below (see, e.g., [19.36] for examples). This is to be expected from the fact that, as discussed in Sect. 19.2.2, this W represents an addition to the Lagrangian and thus to any Hamiltonian, even if only as a boundary term. As for $\Lambda = 0$, the above AAdS arguments were inspired by earlier arguments based on quantum supergravity (see [19.37, 38] for the asymptotically flat case and [19.5] for the AAdS case).

The above paragraph discussed only AAdS spacetimes. While the techniques described there can also be generalized to many AlAdS settings, it is not possible to proceed in this way for truly general choices of M and ∂M. The issue is that the methods of [19.28, 29] require one to find a spinor field satisfying a Dirac-type equation subject to certain boundary conditions. However, for some M, ∂M we can show that no solution exists. In particular, this obstruction arises when $\partial M = S^1 \times \mathbb{R}^{d-1}$ and the S^1 is contractible in M [19.39].

The same obstruction also arises with zero cosmological constant in the context of Kaluza–Klein theories (where the boundary conditions may again involve an S^1 that is contractible in the bulk). In that case, the existence of so-called bubbles of nothing demonstrates that the energy is, in fact, unbounded below and that the system is unstable even in vacuum [19.40, 41]. However, what is interesting about the AlAdS context with $\partial M = S^1 \times \mathbb{R}^{d-1}$ is that there are good reasons [19.39] to believe that the energy *is*, in fact, bounded below – even if there are there are some solutions with energy lower than what one might call empty AdS with $\partial M = S^1 \times \mathbb{R}^{d-1}$ (by which we mean the quotient of the Poincaré patch under some translation of the x^i). Perhaps the strongest such argument (which we will not explain here) comes from AdS/CFT. Another is that [19.42] identified a candidate lowest-energy solution (called the AdS soliton) which was shown [19.39] to at least locally minimize the energy. Proving that the AdS soliton is the true minimum of the energy, or falsifying the conjecture, remains an interesting open problem, whose solution appears to require new techniques.

19.3 Relation to Hamiltonian Charges

We have shown that the charges (19.80) are conserved and motivated their definition in analogy with familiar constructions for field theory in a fixed curved spacetime. It is natural to ask whether the charges (19.80), in fact, agree with more familiar Hamiltonian definitions of asymptotic charges constructed, say, using the AdS generalization of the Hamiltonian approach described in Chap. 17. Denoting these latter charges by $H[\xi]$, the short answer is that they agree as long as we choose s_0 in (19.80) to satisfy $H[\xi](s_0) = 0$; i.e., they agree as long as we choose the same (in principle arbitrary) zero-point for each notion of charge. We may equivalently say that the difference $Q[\xi] - H[\xi]$ is the same for all solutions in our phase space, although for conformal

charges it may depend on the choice of Cauchy surface C for ∂M. As above, for simplicity we take ∂M to be globally hyperbolic with compact Cauchy surfaces.

This result may be found by direct computation (see [19.43] for simple cases). A more elegant, more general, and more enlightening argument can be given [19.10] using a covariant version of the Poisson bracket known as the Peierls bracket [19.44]. The essence of the argument is to show that $Q[\xi]$ generates the canonical transformations associated with the diffeomorphisms ξ. This specifies all Poisson brackets of $Q[\xi]$ to be those of $H[\xi]$. Thus $Q[\xi]-H[\xi]$ must be a c-number in the sense that all Poisson brackets vanish. However, this means that it is constant over the phase space.

After pausing to introduce the Peierls bracket, we sketch this argument below following [19.10]. As in Sect. 19.2.4, we suppose for simplicity that the only bulk fields are the metric and a single scalar field with noninteger ν and we impose boundary conditions that fix both $\gamma_{ij}^{(0)}$ and $\phi^{(0)}$. However, the argument for general bulk fields is quite similar [19.10]. While this material represents a certain aside from our main discussion, it will provide insight into the algebraic properties of conserved charges, the stress tensor itself, and a more general notion of so-called boundary observables that we will shortly discuss.

19.3.1 The Peierls Bracket

The Peierls bracket is a Lie bracket operation that acts on gauge-invariant functions on the space of solutions S of some theory. As shown in the original work [19.44], this operation is equivalent to the Poisson bracket under the natural identification of the phase space with the space of solutions. However, the Peierls bracket is *manifestly* spacetime covariant. In particular, one may directly define the Peierls bracket between any two quantities A and B located anywhere in spacetime, whether or not they may be thought of as lying on the same Cauchy surface. In fact, both A and B can be highly nonlocal, extending over large regions of space and time. These features make the Peierls bracket ideal for studying the boundary stress-tensor, which is well defined on the space of solutions but is not a local function in the bulk spacetime.

To begin, consider two functions A and B on S, which are, in fact, defined as functions on a larger space $\mathcal{H}$, which we call the space of histories. This space $\mathcal{H}$ is the one on which the action is defined; i. e., the solution space S consists of those histories in $\mathcal{H}$ on which the action S is stationary. One may show that the Peierls bracket on S depends only on A, B on S and not on their extensions to $\mathcal{H}$.

The Peierls bracket is defined by considering the effect on one gauge-invariant function (say, B) when the action is deformed by a term proportional to another such function (A). One defines the advanced (D_A^+B) and retarded (D_A^-B) effects of A on B by comparing the original system with a new system given by the action $S_\epsilon = S+\epsilon A$, but associated with the same space of histories $\mathcal{H}$. Here ϵ is a real parameter which will soon be taken to be infinitesimal, and the new action is associated with a new space S_ϵ of deformed solutions.

Under retarded (advanced) boundary conditions for which the solutions $s \in S$ and $s_\epsilon \in S_\epsilon$ coincide in the past (future) of the support of A, the quantity $B_0 = B(s)$ computed using the undeformed solution s will in general differ from $B_\epsilon^\pm = B(s_\epsilon)$ computed using s_ϵ and retarded ($-$) or advanced ($+$) boundary conditions (see Fig. 19.4). For small epsilon, the difference between these quantities defines the retarded (advanced) effect D_A^-B (D_A^+B) of A on B through

$$D_A^\pm B = \lim_{\epsilon \to 0} \frac{1}{\epsilon}(B_\epsilon^\pm - B_0)\,, \tag{19.83}$$

which is a function of the unperturbed solution s. Similarly, one defines $D_B^\pm A$ by reversing the roles of A and B above. Since A, B are gauge invariant, $D_B^\pm A$ is a well-defined (and again gauge-invariant) function on the space S of solutions as long as both A and B are first-differentiable on $\mathcal{H}$. This requirement may be subtle if

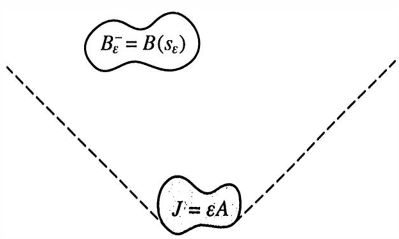


Fig. 19.4 An illustration of the definition of B_ϵ^-. A source term $J=\epsilon A$ is added to the action and the gauge-invariant function B is calculated for the deformed solution s_ϵ subject to the boundary conditions that s and s_ϵ coincide in the far past. *Dashed lines* indicate the boundary of the causal future of J. Only functions B which have support in this region can have $B(s_\epsilon) \neq B(s)$. For visual clarity we have chosen our gauge-invariant function A and B to have compact support, although this is not required

the spacetime supports of A and B extend into the far past and future, but is straightforward for objects like $T^{ij}_{\text{bndy}}(x)$, $\Phi_{\text{bndy}}(x)$ that are well localized in time.

The Peierls bracket [19.44] is then defined to be the difference of the advanced and retarded effects

$$\{A, B\} = D_A^+ B - D_A^- B \,. \tag{19.84}$$

As shown in [19.44], this operation agrees with the Poisson bracket (suitably generalized to allow A, B at unequal times). This generalizes the familiar result that the commutator function for a free scalar field is given by the difference between the advanced and retarded Green's functions. In fact, it is enlightening to write the Peierls bracket more generally in terms of such Green's functions. To do so, let us briefly introduce the notation ϕ^I for a complete set of bulk fields (including the components of the bulk metric) and the associated advanced and retarded Green's functions $G^{\pm}_{IJ}(x, x')$. Note that we have

$$\begin{aligned} D_A^+ B &= \int \mathrm{d}x\, \mathrm{d}x' \frac{\delta B}{\delta \Phi^I(x)} G^+_{IJ}(x, x') \frac{\delta A}{\delta \Phi^J(x')} \\ &= \int \mathrm{d}x\, \mathrm{d}x' \frac{\delta B}{\delta \phi^j(x')} G^-_{JI}(x', x) \frac{\delta A}{\delta \phi^i(x)} \\ &= D_B^- A \,, \end{aligned} \tag{19.85}$$

where we have used the identity $G^+_{IJ}(x,x') = G^-_{JI}(x',x)$. Thus, the Peierls bracket may also be written in the manifestly antisymmetric form

$$\{A, B\} = D_B^- A - D_A^- B = D_A^+ B - D_B^+ A \,. \tag{19.86}$$

The expressions (19.85) in terms of $G^{\pm}_{IJ}(x, x')$ are also useful in order to verify that the Peierls bracket defines a Lie–Poisson algebra. In particular, the derivation property $\{A, BC\} = \{A, B\}C + \{A, C\}B$ follows immediately from the Leibniz rule for functional derivatives. The Jacobi identity also follows by a straightforward calculation, making use of the fact that functional derivatives of the action commute (see, e.g., [19.45, 46]). If one desires, one may use related Green's function techniques to extend the Peierls bracket to a Lie algebra of gauge-dependent quantities [19.47].

19.3.2 Main Argument

We wish to show that the charges $Q[\xi]$ generate the appropriate asymptotic symmetry for any asymptotic Killing field ξ. Since this is true by definition for any Hamiltonian charge $H[\xi]$, it will then follow that $Q[\xi] - H[\xi]$ is constant over the space of solutions S. We first address the case where ξ is compatible with Ω and then proceed to the more general case where $\hat{\xi}$ acts only as a conformal Killing field on the boundary.

Showing that $Q[\xi]$ generates diffeomorphisms along ξ amounts to proving a certain version of Noether's theorem. Recall that the proof of Noether's theorem involves examining the change in the action under a spacetime-dependent generalization of the desired symmetry. The structure of our argument below is similar, where we consider both the action of a given asymptotic symmetry ξ and the spacetime-dependent generalization $f\xi$ defined by choosing an appropriate scalar function f on M. It turns out to be useful to choose f on M (with restriction $\hat{f}$ to ∂M) such that:

- $f = 0$ in the far past and $f = 1$ in the far future.
- $\hat{f} = 0$ to the past of some Cauchy surface C_0 of ∂M, and $\hat{f} = 1$ to the future of some Cauchy surface C_1 of ∂M.

Suppose now that ξ is an asymptotic symmetry compatible with Ω. Then the bulk and boundary fields transform as

$$\begin{aligned} &\delta\phi = \pounds_\xi \phi \,, && \delta g_{\mu\nu} = \pounds_\xi g_{\mu\nu} \,, \\ &\delta\gamma^{(0)}_{ij} = \pounds_{\hat{\xi}} \gamma^{(0)}_{ij} = 0 \,, && \delta\phi^{(0)} = \pounds_{\hat{\xi}} \phi^{(0)} = 0 \,. \end{aligned} \tag{19.87}$$

The key step of the argument is to construct a new transformation $\Delta_{f,\xi}$ on the space of fields such that the associated first-order change $\Delta_{f,\xi} S$ in the action generates the asymptotic symmetry $-\xi$. We will first show that the above property turns out to hold for

$$\Delta_{f,\xi} := (\pounds_{f\xi} - f\pounds_\xi) \,, \tag{19.88}$$

and then verify that $\Delta_{f,\xi} S = -Q[\xi]$. The form of $\Delta_{f,\xi} S$ is essentially that suggested in [19.48] using Hamilton–Jacobi methods, so our argument will also connect $Q[\xi]$ with [19.48].

An important property of (19.88) is that the changes $\Delta_{f,\xi} g_{\mu\nu}$ and $\Delta_{f,\xi} \phi$ are *algebraic* in ϕ and $g_{\mu\nu}$; i.e., we need not take spacetime derivatives of $g_{\mu\nu}, \phi$ to compute the action of $\Delta_{f,\xi}$. Furthermore, $\Delta_{f,\xi}\phi$ and $\Delta_{f,\xi} g_{\mu\nu}$ are both proportional to $\nabla_a f$, and so vanish in both the far future and the far past. This guarantees that $\Delta_{f,\xi} S$ is a differentiable function on $\mathcal{H}$. In particular,

solutions to the equations of motion resulting from the deformed action $S+\epsilon\Delta_{f,\xi}S$ are indeed stationary points of $S+\epsilon\Delta_{f,\xi}S$ under all variations which preserve the conditions and vanish in the far future and past.

It is important to note that the quantity $\Delta_{f,\xi}S$ is gauge-invariant when the equations of motion hold. This is easy to see since by definition on S all variations of S become pure boundary terms. Boundary terms in the far past and future vanish due to the observations above, and since $\gamma_{ij}^{(0)}$, $\phi^{(0)}$ are fixed by boundary conditions the boundary terms on ∂M depend on the bulk fields only through the gauge-invariant quantities T^{ij}_{bndy} and Φ_{bndy}. Thus, we may take the Peierls bracket of $\Delta_{f,\xi}S$ with any other observable A.

We proceed by considering the modified action

$$\begin{aligned}\tilde{S}[\phi,g_{\mu\nu}] &= S[\phi,g_{\mu\nu}]+\epsilon\Delta_{f,\xi}S[\phi,g_{\mu\nu}]\\ &= S[\phi+\epsilon\Delta_{f,\xi}\phi,g_{\mu\nu}+\epsilon\Delta_{f,\xi}g_{\mu\nu}]\,,\end{aligned} \tag{19.89}$$

where the last equality holds to first order in ϵ (and, in fact, defines $\Delta_{f,\xi}S[\phi,g_{\mu\nu}]$). Since $\tilde{S}$ is just S with its argument shifted by $\epsilon\Delta_{f,\xi}$, the stationary points s_1 of $\tilde{S}$ are precisely the oppositely-shifted versions of the stationary points s of S; i. e., we may write $s_1=(1-\epsilon\Delta_{f,\xi})s$ for some $s\in S$.

We should, of course, ask if s_1 satisfies the desired boundary conditions on ∂M. Since ξ is compatible with Ω, the boundary fields shift in the same way as their bulk counterparts; i. e., those of s_1 have been shifted by $-\epsilon\Delta_{f,\xi}$ relative to those of s. Since ξ is an asymptotic symmetry, its action preserves the boundary fields. Now, the reader will note that there is a nontrivial effect from the $\pounds_{f\xi}$ term in $\Delta_{f,\xi}$. This term is a pure diffeomorphism, and since all boundary terms are covariant on ∂M the action $\tilde{S}$ is invariant under *all* diffeomorphisms compatible with Ω (i. e., which preserve the given conformal frame), even those that act nontrivially on the boundary. So the history

$$s_2=(1+\epsilon\pounds_{f\xi})s_1=(1+\epsilon f\pounds_\xi)s \tag{19.90}$$

has

$$\phi^{(0)}|_{s_2}=\phi^{(0)}|_s\,,\quad g_{\mu\nu}|_{s_2}=g_{\mu\nu}|_s\,, \tag{19.91}$$

and again solves the equations of motion that follow from $\tilde{S}$.

This observation allows a straightforward computation of the advanced and retarded changes $D^{\pm}_{\Delta_{f,\xi}S}A$ for any gauge-invariant quantity A. We first consider the retarded change evaluated on a solution s as above. We require a solution s_ϵ^- of the perturbed equations of motion which agrees with s in the far past. Since the infinitesimal transformation $f\pounds_\xi$ vanishes in the far past, we may set $s_\epsilon^-=s_2$ as defined (19.90) above; i. e., $s_\epsilon^-=(1+\epsilon f\pounds_\xi)s$. Thus, the retarded effect on A is just $D^-_{\Delta_{f,\xi}S}A=f\pounds_\xi A$.

To compute the advanced effect, we must find a solution s_ϵ^+ of the perturbed equations of motion which agrees with s in the far future. Consider the history $s_\epsilon^+=(1-\epsilon\pounds_\xi)s_\epsilon^-=(1+(f-1)\epsilon\pounds_\xi)s$. Since this differs from s_ϵ^- by the action of a symmetry compatible with Ω, it again solves the desired equations of motion (to first order in ϵ) and induces the required boundary fields (19.91). In addition, s_ϵ^+ and s agree in the far future (where $f=1$). Thus, we may use s_ϵ^+ to compute the advanced change in any gauge-invariant A

$$D^+_{\Delta_{f,\xi}S}A=(f-1)\pounds_\xi A\,. \tag{19.92}$$

Finally, we arrive at the Peierls bracket

$$\{\Delta_{f,\xi}S,A\}=D^+_{\Delta_{f,\xi}S}A-D^-_{\Delta_{f,\xi}S}A=-\pounds_\xi A\,. \tag{19.93}$$

As desired $-\Delta_{f,\xi}S$ generates a diffeomorphism along the asymptotic symmetry ξ.

All that remains is to relate $\Delta_{f,\xi}S$ to $Q[\xi]$; this is straightforward. Since f vanishes in the far past and future we have

$$\begin{aligned}\Delta_{f,\xi}S &= \int_M\left(\frac{\delta S}{\delta\phi}\Delta_{f,\xi}\phi+\frac{\delta S}{\delta g_{\mu\nu}}\Delta_{f,\xi}g_{\mu\nu}\right)\\ &\quad+\tfrac{1}{2}\int_{\partial M}\sqrt{\gamma^{(0)}}T^{ij}_{\text{bndy}}\Delta_{f,\xi}\gamma^{(0)}_{ij}\\ &\quad+\int_{\partial M}\sqrt{\gamma^{(0)}}\Phi_{\text{bndy}}\Delta_{f,\xi}\phi^{(0)}\,.\end{aligned} \tag{19.94}$$

However, the bulk term vanishes on solutions $s\in S$, and from (19.87) we find $\Delta_{f,\xi}\phi^{(0)}=(\pounds_{\hat{f}\hat{\xi}}-\hat{f}\pounds_{\hat{\xi}})\phi^{(0)}=0$. So only the term containing T^{ij}_{bndy} contributes to (19.94).

To compute the remaining term note that

$$\Delta_{f,\xi}\gamma^{(0)}_{ij}=(\pounds_{\hat{f}\hat{\xi}}-\hat{f}\pounds_{\hat{\xi}})\gamma^{(0)}_{ij}=\hat{\xi}_i\partial_j\hat{f}+\hat{\xi}_j\partial_i\hat{f}\,. \tag{19.95}$$

Since (19.95) vanishes when f is constant, we may restrict the integral over ∂M to the region V between C_0

and C_1 and use the symmetry $T^{ij}_{\text{bndy}} = T^{ji}_{\text{bndy}}$ to obtain

$$\begin{aligned}\Delta_{f,\xi} S &= \int_V \sqrt{|\gamma^{(0)}|} T^{ij}_{\text{bndy}} \xi_i \partial_j f \\ &= \int_{C_1} \sqrt{q} n_j T^{ij}_{\text{bndy}} \xi_i - \int_V \sqrt{|\gamma^{(0)}|} f D_i \left(T^{ij}_{\text{bndy}} \xi_j \right) \\ &= -Q_{C_1}[\xi] \,. \end{aligned} \tag{19.96}$$

Here we used the fact that $\hat{f} = 0$ on C_0 to drop contributions from C_0 and the fact that $\hat{\xi}$ is a Killing field of the boundary metric along with (19.78) to show that the $\int_V$ term in the second line vanishes.

Thus, $-\Delta_{f,\xi} S$ agrees (on solutions) with the charge $Q[\xi]$ evaluated on the cut C_1. Since $Q[\xi]$ is conserved, this equality also holds on any other cut of ∂M. Having already shown by (19.93) that the variation $\Delta_{f,\xi} S$ generates the action of the infinitesimal symmetry $-\xi$ on observables, it follows that $Q[\xi]$ generates the action of ξ

$$\{Q[\xi], A\} = \pounds_\xi A \,, \tag{19.97}$$

as desired.

19.3.3 Asymptotic Symmetries not Compatible with Ω

We now generalize the argument to asymptotic symmetries ξ that are *not* compatible with Ω, so that $\hat{\xi}$ satisfies (19.77). The field content and boundary conditions are the same as above. However, the nontrivial action of ξ on Ω means that there are now are additional terms when a diffeomorphism acts on the boundary fields $\phi^{(0)}, \gamma^{(0)}_{ij}$

$$\begin{aligned}\delta_{\pounds_f \xi} \phi^{(0)} &= \pounds_{\hat{f}\hat{\xi}} \phi^{(0)} - \Delta_- \hat{f} \sigma \phi^{(0)} \,, \\ \delta_{\pounds_f \xi} \gamma^{(0)}_{ij} &= \pounds_{\hat{f}\hat{\xi}} \gamma^{(0)}_{ij} + 2\hat{f} \sigma \gamma^{(0)}_{ij} \,. \end{aligned} \tag{19.98}$$

Combining (19.77) and (19.98) we see that $\delta_{\pounds_\xi}$ acts trivially on the boundary data $\gamma^{(0)}_{ij}, \phi^{(0)}$, as it must since asymptotic symmetries were defined to leave the boundary conditions invariant. Thus the histories $s^{\pm}_\epsilon$ identified above (see, e.g., (19.90)) again satisfy the same boundary conditions as s.

In contrast to Sect. 19.3.2 the operation $\pounds_{f\xi}$ now acts nontrivially on Ω and thus on S. However, since this is only through the conformal anomaly term $a_{(d)}$ in (19.65), $\pounds_{f\xi} S$ depends only on the boundary metric $\gamma^{(0)}$ and is otherwise constant on $\mathcal{H}$. So the equations of motion are unchanged and the histories $s^{\pm}_\epsilon$ again solve the equations of motion for $\tilde{S}$.

It remains to repeat the analog of the calculation (19.96), but here the only change is that the $\int_V$ term on the second line no longer vanishes. Instead, it contributes a term proportional to $a_{(d)}$. Since this term is constant on the space of solutions S, it has vanishing Peierls brackets, and we again conclude that $Q_{C_1}[\xi]$ generates the asymptotic symmetry ξ. (This comment corrects a minor error in [19.47].) Moreover, since $Q_C[\xi]$ depends on the Cauchy surface C only through a term that is constant on S, the same result holds for any C. Thus, even when $\hat{\xi}$ is only a conformal symmetry of the boundary, $Q_C[\xi] - H[\xi]$ is constant over the space S of solutions.

19.3.4 Charge Algebras and Central Charges

We saw above that our charges $Q[\xi]$ generate the desired asymptotic symmetries via the Peierls bracket. This immediately implies what is often called the *representation theorem* that the algebra of the charges themselves matches that of the associated symmetries up to possible so-called central extensions. This point is really quite simple. Consider three vector fields ξ_1, ξ_2, ξ_3 related via the Lie bracket through $\{\xi_1, \xi_2\} = \xi_3$. Now examine the Jacobi identity

$$\begin{aligned}&\{Q[\xi_1], \{Q[\xi_2], A\}\} + \{Q[\xi_2], \{A, Q[\xi_1]\}\} \\ &+ \{A, \{Q[\xi_1], Q[\xi_2]\}\} = 0 \,, \end{aligned} \tag{19.99}$$

which must hold for any A. Since $\{Q[\xi_i], B\} = \pounds_{\xi_i} B$ for any B, we may use (19.99) to write

$$\begin{aligned}\pounds_{\xi_3} A &= \pounds_{\xi_1}(\pounds_{\xi_2} A) - \pounds_{\xi_2}(\pounds_{\xi_1} A) \\ &= \{\{Q[\xi_2], Q[\xi_1]\}, A\} \,. \end{aligned} \tag{19.100}$$

The left-hand-side is also $\{Q[\xi_3], A\}$, so we conclude that $\{Q[\xi_1], Q[\xi_2]\}$ generates the same transformation as $Q[\xi_3]$. This means that they can differ only by some $K(\xi_1, \xi_2)$, which is constant across the space of solutions (i. e., it is a so-called c-number)

$$\{Q[\xi_1], Q[\xi_2]\} = Q[\{\xi_1, \xi_2\}] + K(\xi_1, \xi_2) \,. \tag{19.101}$$

For some symmetry algebras one can show that any such $K(\xi_i, \xi_j)$ can be removed by shifting the zero-points of the charges $Q[\xi]$. In such cases the $K(\xi_i, \xi_j)$ are

said to be trivial. Nontrivial $K(\xi_i, \xi_j)$ are classified by a cohomology problem and are said to represent central extensions of the symmetry algebra.

It is easy to show that $K(\xi_i, \xi_j)$ may be set to zero in this way whenever there is some solution (call it s_0) which is invariant under all symmetries. The fact that it is invariant means that $\{Q[\xi_i], A\}(s_0) = 0$; i.e., the bracket vanishes when evaluated on the particular solution s_0 for any ξ_i and any A. So take $A = Q[\xi_j]$, and set the zero-points of the charges so that $Q[\xi](s_0) = 0$. Evaluating (19.101) on s_0 then gives $K(\xi_i, \xi_j)(s_0) = 0$ for all ξ. However, since $K(\xi_i, \xi_j)(s_0)$ is constant over the space of solutions, this means that it vanishes identically.

For asymptotically flat spacetimes the asymptotic symmetries generate the Poincaré group, which are just the exact symmetries of Minkowski space. Thus one might expect the asymptotic symmetries of $(d+1)$-dimensional AlAdS spacetimes to be (perhaps a subgroup of) $SO(d,2)$ in agreement with the isometries of AdS_{d+1} compatible with the boundary conditions on ∂M. Since (at least when it is allowed by the boundary conditions) empty AdS_{d+1} is a solution invariant under all symmetries, one might expect that the corresponding central extensions are trivial.

This turns out to be true for $d > 2$. Indeed, any Killing field of AdS_{d+1} automatically satisfies our definition of an asymptotic symmetry (at least for boundary conditions $\phi^{(0)} = 0$ and $\gamma_{ij}^{(0)}$ the metric on the Einstein static universe). However, for $d = 2$ there are additional asymptotic Killing fields that are not Killing fields of empty AdS_3. This is because all $d = 2$ boundary metrics $\gamma_{ij}^{(0)}$ take the form $\mathrm{d}s^2 = g_{uv}\,\mathrm{d}u\,\mathrm{d}v$ when written in terms of null coordinates, making manifest that any vector field $\hat{\xi}^u = f(u)$, $\hat{\xi}^v = g(v)$ is a conformal Killing field of $\gamma_{ij}^{(0)}$. This leads to an infinite-dimensional asymptotic symmetry group, which is clearly much larger than the group $SO(2,2)$ of isometries of AdS_3.

Thus as first noted in [19.8] there can be a nontrivial central extension for $d = 2$. In this case, one can show that up to the above-mentioned zero-point shifts all central extensions are parametrized by a single number c called the central charge. (When parity symmetry is broken, there can be separate left and right central charges c_L, c_R.) Reference [19.8] calculated this central charge using Hamiltonian methods, but we will follow [19.26] and work directly with the boundary stress tensor.

Since the charges $Q[\xi]$ generate (bulk) diffeomorphisms along ξ, and since the charges themselves are built from T^{ij}_{bndy}, the entire effect is captured by computing the action of a bulk diffeomorphism ξ on T^{ij}_{bndy}. As noted in Sect. 19.1.5, the action of ξ on boundary quantities generally involves both a diffeomorphism $\hat{\xi}$ along the boundary and a change of conformal frame. Moreover, as we have seen, for even d changes of conformal frame act nontrivially on T^{ij}_{bndy}. For $g_{uv} = -1$ a direct calculation gives

$$
\begin{aligned}
T_{\mathrm{bndy}\,uu} \to\ & T_{\mathrm{bndy}\,uu} \\
& + \left(2T_{\mathrm{bndy}\,uu}\partial_u\xi^u + \xi^u\partial_u T_{\mathrm{bndy}\,uu}\right) \\
& - \frac{c}{24\pi}\partial_u^3\xi^u \\
T_{\mathrm{bndy}\,vv} \to\ & T_{\mathrm{bndy}\,vv} \\
& + \left(2T_{\mathrm{bndy}\,vv}\partial_v\xi^v + \xi^v\partial_v T_{\mathrm{bndy}\,vv}\right) \\
& - \frac{c}{24\pi}\partial_v^3\xi^v\,,
\end{aligned}
\tag{19.102}
$$

where $c = 3\ell/2G$. The term in parenthesis is the tensorial part of the transformation, while the final (so-called anomalous) term is associated with the conformal anomaly $a_{(2)} = -(c/24\pi)\mathcal{R}$.

It is traditional to Fourier transform the above components of the stress tensor to write the charge algebra as the (double) Virasoro algebra

$$
\mathrm{i}\{L_m, L_n\} = (m-n)L_{m+n} + \frac{c}{12}m(m^2-1)\delta_{m+n,0}\,,
\tag{19.103}
$$

$$
\mathrm{i}\{\bar{L}_m, \bar{L}_n\} = (m-n)\bar{L}_{m+n} + \frac{c}{12}m(m^2-1)\delta_{m+n,0}\,,
\tag{19.104}
$$

where $\{L_n, \bar{L}_m\} = 0$ and

$$
\begin{aligned}
L_n &= -\frac{1}{2\pi}\int_{S^1} \mathrm{e}^{\mathrm{i}un}T_{\mathrm{bndy}\,uu}\,\mathrm{d}u\,, \\
\bar{L}_n &= -\frac{1}{2\pi}\int_{S^1} \mathrm{e}^{\mathrm{i}vn}T_{\mathrm{bndy}\,vv}\,\mathrm{d}v\,.
\end{aligned}
\tag{19.105}
$$

Here we have take $\partial M = S^1 \times \mathbb{R}$ so that the dynamics requires both T_{uu} and T_{vv} to be periodic functions of their arguments. We have taken this period to be 2π.

The anomalous transformation of T^{ij}_{bndy} leads to interesting zero-points for certain charges. Suppose, for example, that we take T^{ij}_{bndy} to vanish for the Poincaré patch of empty AdS_3 in the conformal frame where the boundary metric is (uncompactified) Minkowski space.

Then, since $S^1 \times \mathbb{R}$ is (locally) conformal to Minkowski space, we can use the conformal anomaly to calculate T^{ij}_{bndy} for empty AdS_3 with the Einstein static universe boundary metric. One finds that the resulting energy does not vanish. Instead, $E_{\text{global AdS}_3} = -c/12\ell = -1/8G$ so that $E = 0$ for the so-called $M = 0$ Bañados-Teitelboim–Zanelli (BTZ) black hole [19.49, 50]. The offset in (19.82) arises from similarly setting $T^{ij}_{\text{bndy}} = 0$ for empty AdS_5 in the conformal frame where the boundary metric is (uncompactified) Minkowski space.

19.4 The Algebra of Boundary Observables and the AdS/CFT Correspondence

We have shown above how the boundary stress tensor can be used to construct charges $Q[\xi]$ associated with any asymptotic symmetry ξ of a theory of asymptotically locally AdS spacetimes. The $Q[\xi]$ are conserved (perhaps, up to c-number anomaly terms) and generate the asymptotic symmetry ξ under the action of the Peierls bracket (or equivalently, under the Poisson bracket). Therefore, $Q[\xi]$ are equivalent to the Hamiltonian charges that we could derive using techniques analogous to those described in Chap. 17 for asymptotically flat spacetimes. Conversely, boundary stress tensor methods can also be applied in the asymptotically flat context [19.51–53]. Readers interested in direct Hamiltonian approaches to AdS charges should consult [19.6–8]; see also [19.5, 13, 14, 54–57] for other covariant approaches.

We chose to use boundary stress tensor methods for two closely related reasons. The first is that, in addition to its role in constructing conserved charges, the local boundary field T^{ij}_{bndy} turns out to contain useful information on its own. For example, it plays a key role in the hydrodynamic description of large AdS black holes known as the fluid/gravity correspondence [19.58] (which may be considered a modern incarnation of the so-called membrane paradigm [19.59]). The extra information in T^{ij}_{bndy} appears at the AdS boundary ∂M due to the fact that all multipole moments of a given field decay near ∂M with the same power law; namely, the one given by the $\gamma^{(d)}$ term in the Fefferman–Graham expansion (19.21). This is in striking contrast with the more familiar situation in asymptotically flat spacetimes where the large r behavior is dominated by the monopole terms, with subleading corrections from the dipole and higher order multipoles. Indeed, while as noted above similar boundary stress tensor techniques can be employed in asymptotically flat spacetimes, the asymptotically flat boundary stress tensor contains far less information.

The second reason is that both T^{ij}_{bndy} and Φ_{bndy} play fundamental roles in the AdS/CFT correspondence [19.2] (see especially [19.4]). Any treatment of asymptotic AdS charges would be remiss without at least mentioning this connection, and we take the opportunity below to give a brief introduction to AdS/CFT from the gravity side. This turns out to be straightforward using the machinery described thus far. Indeed, the general framework requires no further input from either string theory or CFT and should be readily accessible to all readers of this volume. As usual, we consider bulk gravity coupled to a single bulk scalar and fix both $\gamma^{(0)}_{ij}$ and $\phi^{(0)}$ as boundary conditions. We refer to $\gamma^{(0)}_{ij}$ and $\phi^{(0)}$ as boundary sources below. More general boundary conditions may be thought of as being dual to CFTs with additional interactions [19.60] or coupled to additional dynamical fields [19.27, 61, 62], although we will not go into the details here.

The only new concept we require is that of the algebra $\mathcal{A}_{\text{bndy}}$ of boundary observables, which is just the algebra generated by T^{ij}_{bndy} and Φ_{bndy} under the Peierls bracket. Here we mean that we consider the smallest algebra containing both T^{ij}_{bndy} and Φ_{bndy}, which is closed under finite flows; i. e., under the classical analog of the quantum operation $e^{iA}Be^{-iA}$. A key property of $\mathcal{A}_{\text{bndy}}$ follows from the fact that the bulk equations of motion are completely independent of the choice of conformal frame Ω. Thus, up to the usual conformal anomalies, under any change of conformal frame the boundary observables transform only by rescaling with a particular power of $e^{-\sigma}$ known as the conformal dimension (d for T^{ij}_{bndy}, and Δ_+ for Φ_{bndy}), with the boundary sources transforming similarly with conformal weights zero for $\gamma^{(0)}_{ij}$ and Δ_- for $\phi^{(0)}$. (In defining the conformal dimension it is conventional not to count the ± 2 powers of $e^{-\sigma}$ associated with the indices on T^{ij}_{bndy} and $\gamma^{(0)}_{ij}$.) In this sense the theory of $\mathcal{A}_{\text{bndy}}$ is invariant (or, perhaps better, covariant) under all changes of boundary conformal frame. Of course we have already shown that when the boundary observables admit a conformal Killing field $\hat{\xi}$, the corresponding transformation is generated by the associated $Q[\xi]$ from (19.79). Now

since the charges $Q[\xi]$ are built from T^{ij}_{bndy} and Φ_{bndy} they also lie in the algebra $\mathcal{A}_{\text{bndy}}$. When ξ can be chosen to be everywhere time like, this immediately implies that $\mathcal{A}_{\text{bndy}}$ is also closed under time evolution. This last property can also be shown much more generally; see, e.g., [19.63].

We now extract one final property of the algebra $\mathcal{A}_{\text{bndy}}$. From the expression (19.85) in terms of Green's functions, it is clear that the Peierls bracket $\{A, B\}$ of two observables vanishes on any solution s for which A, B are outside each other's light cones; i. e., when the regions on which A, B are supported cannot be connected by any causal curve. Furthermore, as shown in [19.64] the null energy condition implies that two boundary points x, y can be connected by a causal curve through the bulk only when they can also be connected by a causal curve lying entirely in the boundary. It follows that the algebra $\mathcal{A}_{\text{bndy}}$ satisfies the usual definition of locality for a field theory on ∂M; namely that Peierls brackets vanish outside the light cones defined by the boundary metric $\gamma^{(0)}_{ij}$.

Although we have so far worked entirely at the classical level, let us now assume that all of the above properties persist in the quantum theory. We then have a conformally covariant algebra of operators $\mathcal{A}_{\text{bndy}}$ with closed dynamics, local commutation relations on ∂M, and a stress tensor T^{ij}_{bndy} that generates all conformal symmetries. In other words, we have a local CFT on ∂M.

This is the most basic statement of the AdS/CFT correspondence. *Any* bulk AlAdS quantum gravity theory in which the above classical properties continue to hold defines a CFT through its algebra $\mathcal{A}_{\text{bndy}}$ of boundary observables. Now, we should remark that the AdS/CFT correspondence as used in string theory goes one step further. For certain specific bulk theories it identifies the so-called dual CFT as a particular known theory defined by its own Lagrangian with a definite field content. For example, when the bulk is type IIB string theory [19.65] asymptotic to a certain $\text{AdS}_5 \times S^5$ solution, the corresponding CFT is just $\mathcal{N} = 4$ super-Yang–Mills. We will not go into further details here, although the interested reader may consult various reviews such as [19.66–68].

On the other hand, even without having a separate definition of the CFT, the above observations already have dramatic implications for the bulk quantum gravity theory. In particular, the statement that $\mathcal{A}_{\text{bndy}}$ is closed under time evolution runs completely counter to one's usual intuition regarding field theory with a boundary. We usually think that most of the dynamical degrees of freedom live in the bulk spacetime, with perhaps only a small subset visible on the boundary at any time. In particular, we expect any signal present on the boundary at time t_0 to then propagate into the bulk and (at least for some time) to essentially disappear from the algebra of boundary observables. Since $\mathcal{A}_{\text{bndy}}$ is closed under time evolution, it is clear that this is simply not the case in our quantum gravity theory. The difference arises precisely from the fact that the gravitational Hamiltonian (and more generally any $Q[\xi]$) is a pure boundary term. This property was called *boundary unitarity* in [19.63]. See also [19.69] for further discussion of this point.

The reader should take care to separate boundary unitarity from the possible claim that $\mathcal{A}_{\text{bndy}}$ captures the complete set of bulk observables. The two ideas are logically separate, as there can, in principle, be additional bulk observables $\mathcal{A}_{\text{other}}$ as long as they do not mix dynamically with those in $\mathcal{A}_{\text{bndy}}$. One says that the possible values of $\mathcal{A}_{\text{other}}$ define superselection sectors with respect to $\mathcal{A}_{\text{bndy}}$ [19.70]. However, any such additional observables are clearly very special. The requirement that they not affect $\mathcal{A}_{\text{bndy}}$ strongly suggests that at least semi-classically such observables have to do only with properties of spacetime hidden from the boundary behind both past and future horizons [19.71]. In particular, any degrees of freedom that determine whether black holes are connected by (nonreversible) wormholes seem likely to lie in $\mathcal{A}_{\text{other}}$. On the other hand, in perturbation theory about empty AdS (or even about solutions that are empty AdS in the far past) one may show that $\mathcal{A}_{\text{other}}$ is, indeed, empty [19.63].

References

19.1 G.T. Horowitz, V.E. Hubeny: CFT description of small objects in AdS, J. High Energy Phys. **0010**, 027 (2000)

19.2 J.M. Maldacena: The large *N* limit of superconformal field theories and supergravity, Adv. Theor. Math. Phys. **2**, 231–252 (1998)

19.3 S.S. Gubser, I.R. Klebanov, A.M. Polyakov: Gauge theory correlators from noncritical string theory, Phys. Lett. B **428**, 105–114 (1998)

19.4 E. Witten: Anti-de Sitter space and holography, Adv. Theor. Math. Phys. **2**, 253–291 (1998)

19.5 L.F. Abbott, S. Deser: Stability of gravity with a cosmological constant, Nucl. Phys. B **195**, 76 (1982)

19.6 M. Henneaux, C. Teitelboim: Hamiltonian treatment of asymptotically anti-de Sitter spaces, Phys. Lett. B **142**, 355–358 (1984)

19.7 M. Henneaux, C. Teitelboim: Asymptotically anti-de Sitter spaces, Commun. Math. Phys. **98**, 391–424 (1985)
19.8 J.D. Brown, M. Henneaux: Central charges in the canonical realization of asymptotic symmetries, Commun. Math. Phys. **104**, 207–226 (1986)
19.9 I. Papadimitriou, K. Skenderis: Thermodynamics of asymptotically locally AdS spacetimes, J. High Energy Phys. **0508**, 004 (2005)
19.10 S. Hollands, A. Ishibashi, D. Marolf: Counter-term charges generate bulk symmetries, Phys. Rev. D **72**, 104025 (2005)
19.11 T. Andrade, D. Marolf: AdS/CFT beyond the unitarity bound, J. High Energy Phys. **1201**, 049 (2012)
19.12 R.M. Wald: *General Relativity*, 1st edn. (University of Chicago Press, Chicago 1984)
19.13 A. Ashtekar, A. Magnon: Asymptotically anti-de Sitter space-times, Class. Quantum Gravity **1**, L39–L44 (1984)
19.14 A. Ashtekar, S. Das: Asymptotically anti-de Sitter space-times: Conserved quantities, Class. Quantum Gravity **17**, L17–L30 (2000)
19.15 M.C.N. Cheng, K. Skenderis: Positivity of energy for asymptotically locally AdS spacetimes, J. High Energy Phys. **0508**, 107 (2005)
19.16 K. Skenderis: Asymptotically anti-de Sitter spacetimes and their stress energy tensor, Int. J. Mod. Phys. A **16**, 740–749 (2001)
19.17 K. Skenderis: Lecture notes on holographic renormalization, Class. Quantum Gravity **19**, 5849–5876 (2002)
19.18 S. de Haro, S.N. Solodukhin, K. Skenderis: Holographic reconstruction of space-time and renormalization in the AdS/CFT, Commun. Math. Phys. **217**, 595–622 (2001)
19.19 R. Penrose, W. Rindler: *Spinors and Space-Time*, Vol. 2 (Cambridge Univ. Press, Cambridge 1984)
19.20 C. Fefferman, C.R. Graham: Conformal invariants. In: *Elie Cartan et les Mathématiques d'aujourd'hui*, (Société Mathématique de France, Paris 1985) pp. 95–116
19.21 K. Skenderis, S.N. Solodukhin: Quantum effective action from the AdS/CFT correspondence, Phys. Lett. B **472**, 316–322 (2000)
19.22 P. Breitenlohner, D.Z. Freedman: Positive energy in anti-de Sitter backgrounds and gauged extended supergravity, Phys. Lett. B **115**(3), 197–201 (1982)
19.23 J.D.J.W.Y. Brown Jr.: Quasilocal energy and conserved charges derived from the gravitational, Phys. Rev. D **47**, 1407–1419 (1993)
19.24 S.W. Hawking: The path-integral approach to quantum gravity. In: *General Relativity: An Einstein Centenary Survey*, ed. by S.W. Hawking, W. Israel (Cambridge Univ. Press, Cambridge 1979) pp. 746–789
19.25 M. Henningson, K. Skenderis: The holographic Weyl anomaly, J. High Energy Phys. **9807**, 023 (1998)
19.26 V. Balasubramanian, P. Kraus: A Stress tensor for Anti-de Sitter gravity, Commun. Math. Phys. **208**, 413–428 (1999)
19.27 G. Compere, D. Marolf: Setting the boundary free in AdS/CFT, Class. Quantum Gravity **25**, 195014 (2008)
19.28 E. Witten: A simple proof of the positive energy theorem, Commun. Math. Phys. **80**, 381 (1981)
19.29 J.A. Nester: A new gravitational energy expression with a simple positivity proof, Phys. Lett. A **83**, 241 (1981)
19.30 P.K. Townsend: Positive energy and the scalar potential in higher dimensional (super)gravity, Phys. Lett. B **148**, 55 (1984)
19.31 T. Hertog, S. Hollands: Stability in designer gravity, Class. Quantum Gravity **22**, 5323–5342 (2005)
19.32 A.J. Amsel, D. Marolf: Energy bounds in designer gravity, Phys. Rev. D **74**, 064006 (2006)
19.33 A.J. Amsel, T. Hertog, S. Hollands, D. Marolf: A Tale of two superpotentials: Stability and instability in designer, Phys. Rev. D **75**, 084008 (2007)
19.34 T. Faulkner, G.T. Horowitz, M.M. Roberts: New stability results for Einstein scalar gravity, Class. Quantum Gravity **27**, 205007 (2010)
19.35 A.J. Amsel, M.M. Roberts: Stability in Einstein-scalar gravity with a logarithmic branch, Phys. Rev. D **85**, 106011 (2012)
19.36 T. Hertog: Violation of energy bounds in designer gravity, Class. Quantum Gravity **24**, 141–154 (2007)
19.37 S. Deser, C. Teitelboim: Supergravity has positive energy, Phys. Rev. Lett. **39**, 249 (1977)
19.38 M.T. Grisaru: Positivity of the energy in Einstein theory, Phys. Lett. B **73**, 207 (1978)
19.39 G.T. Horowitz, R.C. Myers: The AdS/CFT correspondence and a new positive energy conjecture for general, Phys. Rev. D **59**, 026005 (1998)
19.40 E. Witten: Instability of the Kaluza–Klein vacuum, Nucl. Phys. B **195**, 481 (1982)
19.41 D. Brill, H. Pfister: States of negative total energy in Kaluza–Klein theory, Phys. Lett. B **228**, 359–362 (1989)
19.42 E. Witten: Anti-de Sitter space, thermal phase transition, and confinement in gauge, Adv. Theor. Math. Phys. **2**, 505–532 (1998)
19.43 S. Hollands, A. Ishibashi, D. Marolf: Comparison between various notions of conserved charges in asymptotically, Class. Quantum Gravity **22**, 2881–2920 (2005)
19.44 R.E. Peierls: The commutation laws of relativistic field theory, Proc. R. Soc. **214**, 143 (1952)
19.45 B.S. DeWitt: *Dynamical Theory of Groups and Fields* (Gordon and Breach, Philadelphia 1965)
19.46 B.S. DeWitt: The spacetime approach to quantum field theory. In: *Relativity, Groups, and Topology II: Les Houches. Part 2*, ed. by B.S. Dewitt, R. Stora (North-Holland, Amsterdam 1984)
19.47 D.M. Marolf: The generalized Peierls bracket, Anna. Phys. **236**, 392–412 (1994)

19.48 R.D. Sorkin: Conserved Quantities as Action Variations, Mathematics and General Relativity. Proc. AMS-IMS-SIAM Joint Summer Res. Conf. Santa Cruz, California (Amer. Math. Soc., Providence 1986) pp. 23–37
19.49 M. 8anados, C. Teitelboim, J. Zanelli: The black hole in three-dimensional space-time, Phys. Rev. Lett. **69**, 1849–1851 (1992)
19.50 M. Banados, M. Henneaux, C. Teitelboim, J. Zanelli: Geometry of the $(2+1)$ black hole, Phys. Rev. D **48**, 1506–1525 (1993)
19.51 R.B. Mann, D. Marolf: Holographic renormalization of asymptotically flat spacetimes, Class. Quantum Gravity **23**, 2927–2950 (2006)
19.52 R.B. Mann, D. Marolf, A. Virmani: Covariant counterterms and conserved charges in asymptotically flat, Class. Quantum Gravity **23**, 6357–6378 (2006)
19.53 R.B. Mann, D. Marolf, R. McNees, A. Virmani: On the stress tensor for asymptotically flat gravity, Class. Quantum Gravity **25**, 225019 (2008)
19.54 G.W. Gibbons, S.W. Hawking, G.T. Horowitz, M.J. Perry: Positive mass theorems for black holes, Commun. Math. Phys. **88**, 295 (1983)
19.55 G.W. Gibbons, C.M. Hull, N.P. Warner: The stability of gauged supergravity, Nucl. Phys. B **218**, 173 (1983)
19.56 J. Katz, J. Bicak, D. Lynden-Bell: Relativistic conservation laws and integral constraints for large, Phys. Rev. D **55**, 5957–5969 (1997)
19.57 N. Deruelle, J. Katz: On the mass of a Kerr-anti-de Sitter spacetime in D dimensions, Class. Quantum Gravity **22**, 421–424 (2005)
19.58 S. Bhattacharyya, V.E. Hubeny, S. Minwalla, M. Rangamani: Nonlinear fluid dynamics from gravity, J. High Energy Phys. **0802**, 045 (2008)
19.59 K.S. Thorne, R.H. Price, D.A. Macdonald: *Black Holes: The Membrane Paradigm* (Yale Univ. Press, New Haven 1986)
19.60 E. Witten: Multitrace operators, boundary conditions, and AdS/CFT correspondence (2001), hep-th/0112258
19.61 S.S. Gubser, I. Mitra: Double trace operators and one loop vacuum energy in AdS/CFT, Phys. Rev. D **67**, 064018 (2003)
19.62 E. Witten: SL(2,Z) action on three-dimensional conformal field theories with Abelian (2003), hep-th/0307041
19.63 D. Marolf: Unitarity and holography in gravitational physics, Phys. Rev. D **79**, 044010 (2009)
19.64 S. Gao, R.M. Wald: Theorems on gravitational time delay and related issues, Class. Quantum Gravity **17**, 4999–5008 (2000)
19.65 J. Polchinski: *String Theory*, 1st edn. (Cambridge Univ. Press, Cambridge 1984)
19.66 O. Aharony, S.S. Gubser, J.M. Maldacena, H. Ooguri, Y. Oz: Large *N* field theories, string theory and gravity, Phys. Rept. **323**, 183–386 (2000)
19.67 E. D'Hoker, D.Z. Freedman: Supersymmetric gauge theories and the AdS/CFT correspondence (2002), hep-th/0201253
19.68 J. Polchinski: Introduction to gauge/gravity duality (2010), arXiv:1010.6134
19.69 D. Marolf: Holographic thought experiments, Phys. Rev. D **79**, 024029 (2009)
19.70 D. Marolf: Black holes, AdS, and CFTs, Gen. Rel. Grav. **41**, 903–917 (2009)
19.71 D. Marolf, A.C. Wall: Eternal black holes and superselection in AdS/CFT (2012), arXiv:1210.3590

20. Spacetime Singularities

Pankaj S. Joshi

We give here an overview of our basic understanding of and recent developments in spacetime singularities in the Einstein theory of gravity. Several issues related to physical significance and implications of singularities are discussed. The nature and existence of singularities are considered which indicate the formation of super ultra-dense regions in the universe as predicted by the general theory of relativity. Such singularities develop during the gravitational collapse of massive stars and in cosmology at the origin of the universe. Possible astrophysical implications of the occurrence of singularities in the spacetime universe are indicated. We discuss in some detail the profound and key fundamental issues that the singularities give rise to, such as the cosmic censorship and predictability in the universe, naked singularities in gravitational collapse and their relevance in black hole physics today, and their astrophysical implications in modern relativistic astrophysics and cosmology.

20.1 Space, Time and Matter

After Einstein proposed the general theory of relativity describing the gravitational force in terms of spacetime curvatures, the proposed field equations related the spacetime geometry to the matter content of the universe. The early solutions found for these equations were the Schwarzschild metric and the Friedmann models. While the first represented the gravitational field around an isolated body such as a spherically symmetric star, the later solutions described the geometry of the universe. Both these models contained a spacetime singularity where the curvatures as well as the matter and energy densities diverged and became arbitrarily high, and the physical description would then break down. In the Schwarzschild solution such a singularity was present at the center of symmetry $r = 0$, whereas for the Friedmann models it was found at the epoch $t = 0$,

which is the beginning of the universe and origin of time where the scale factor for the universe vanishes and all objects are crushed to a zero volume due to infinite gravitational tidal forces.

Even though the physical problem posed by the existence of such a strong curvature singularity was realized immediately in these solutions, which turned out to have several important implications for the experimental verification of the general relativity theory, initially this phenomenon was not taken seriously. It was generally thought that the existence of such a singularity must be a consequence of the very high degree of symmetry imposed on the spacetime while these solutions were being derived and obtained. Subsequently, the distinction between a genuine spacetime singularity and a mere coordinate singularity became clear and it was realized that the singularity at $r = 2m$ in the Schwarzschild spacetime was only a coordinate singularity which could be removed by a suitable coordinate transformation. It was clear, however, that the genuine curvature singularity at $r = 0$ cannot be removed by any coordinate transformations. The hope was then that when more general solutions to the field equations are considered with a lesser degree of symmetry, such singularities will be avoided.

This issue was sorted out when a detailed study of the structure of a general spacetime and the associated problem of singularities was taken up by Hawking, Penrose, and Geroch (see, for example, [20.1] and references therein). It was shown by this work that a spacetime will admit singularities within a rather general framework provided that it satisfies certain reasonable physical assumptions such as the positivity of energy, a suitable causality assumption and a condition implying strong gravitational fields, such as the existence of trapped surfaces. It thus followed that the spacetime singularities form a general feature of the relativity theory. In fact, these considerations ensure the existence of singularities in other theories of gravity, which are also based on a spacetime manifold framework and satisfy the general conditions such as those stated above.

Therefore, the scenario that emerges is the following: essentially for all classical spacetime theories of gravitation, the occurrence of singularities forms an inevitable and integral part of the description of the physical reality. In the vicinity of such a singularity, typically the energy densities, spacetime curvatures, and all other physical quantities would blow up, thus indicating the occurrence of super ultra-dense regions in the universe. The behavior of such regions may not be governed by the classical theory itself, which may break down having predicted the existence of the singularities, and a quantum gravitation theory would be the most likely description of the phenomena created by such singularities.

Further to the general relativity theory in 1915, gravitation physics was a relatively quiet field with few developments until about the 1950s. However, the 1960s saw the emergence of new observations in high energy astrophysics, such as quasars and high energy phenomena at the center of galaxies such as extremely energetic jets. These observations, together with important theoretical developments such as studying the global structure of spacetimes and singularities, led to important results in black hole physics and relativistic astrophysics and cosmology.

Our purpose here is to review some of these rather interesting as well as intriguing developments, in a somewhat pedagogic and elementary fashion to provide a broad perspective. Specifically, we would like to highlight here several recent issues and challenges that have emerged related to spacetime singularities, which appear to have a considerable physical significance and may have interesting astrophysical implications. We take up and consider here important topics such as what is meant by a singular spacetime and specify the notion of a singularity. It turns out that it is the notion of geodesic incompleteness that characterizes a singularity in an effective manner for a spacetime and enables their existence to be proved by means of certain relatively general theorems. We highlight here several recent developments which deal with certain exciting current issues related to spacetime singularities on which research in gravitation and cosmology is being carried out today. These include the work on final end states of gravitational collapse, cosmic censorship, and black holes and naked singularities. Related major cosmic conundrums such as the issue of predictability in the universe are discussed and observational implications of naked singularities are indicated.

20.2 What Is a Singularity?

In the Einstein theory of gravitation, the universe is modeled as a spacetime with a mathematical structure of a four-dimensional differentiable manifold. In that case, locally the spacetime is always flat in a sufficiently small region around any point, but on a larger scale this need not be the case and it can have a rich and varied structure. An example of a differentiable manifold is a sphere, which is flat enough in the vicinity of any single point on its surface, but has a global curvature. For a more detailed discussion on spacetime manifolds and their key role in general relativity, we refer to [20.2].

When should we say that a spacetime universe (M, g), which is a differentiable manifold with a Lorentzian metric, has become singular? What we need is a specification and a specific criterion for the existence of a singularity for any given universe model in general relativity.

As stated above, several examples of singular behavior in spacetime models of general relativity are known. Important exact solutions such as the Friedmann–Robertson–Walker (FRW) cosmological models and Schwarzschild spacetime contain a singularity where the energy density or curvatures diverge strongly and the usual description of the spacetime breaks down. In the Schwarzschild spacetime there is an essential curvature singularity at $r=0$ in that along any nonspace-like trajectory falling into the singularity, the Kretschmann scalar $\alpha = R^{ijkl}R_{ijkl} \to \infty$. Also, all future directed nonspace-like geodesics which enter the event horizon at $r = 2m$ must fall into this curvature singularity within a finite value of the proper time, or finite value of the affine parameter, so all such curves are future geodesically incomplete. Similarly, for FRW models, if $\rho + 3p > 0$ at all times, where ρ is the total energy density and p the pressure, there is a singularity at $t = 0$ which could be identified as the origin of the universe. Then along all the past directed trajectories meeting this singularity, $\rho \to \infty$ and the curvature scalar $R = R_{ij}R^{ij} \to \infty$. Again, all the past directed nonspace-like geodesics are incomplete. This essential singularity at $t = 0$ cannot be transformed away by any coordinate transformations. Similar behavior was generalized to the class of spatially homogeneous cosmological models by *Ellis* and *King* [20.3] which satisfy the positivity of energy conditions.

Such singularities, where the curvature scalars and densities diverge imply a genuine spacetime pathology where the usual laws of physics break down. The existence of the geodesic incompleteness in these spacetimes implies, for example, that a time-like observer suddenly disappears from the spacetime after a finite amount of proper time. Of course, singular behavior can also occur without bad behavior of curvature. For example, consider Minkowski spacetime with a point deleted. Then there will be time-like geodesics running into the hole which will be future incomplete. Clearly, this is an artificial situation one would like to rule out by requiring that the spacetime is *inextendible*, that is, it cannot be isometrically embedded into another larger spacetime as a proper subset. However, one could give a nontrivial example of singular behavior where a conical singularity exists (see, e.g., [20.4]). Here spacetime is inextendible but curvature components do not diverge near the singularity, as in a Weyl-type solution. The metric is given by

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + \mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\, \mathrm{d}\phi^2)\,,$$

with coordinates given by $-\infty < t < \infty, 0 < r < \infty, 0 < \theta < \pi$ with $0 < \phi < a$, with $\phi = 0$ and $\phi = a$ identified and $a \neq 2\pi$. There is a conical singularity at $r = 0$ through which the spacetime cannot be extended and the singular boundary is related to the time-like two-plane $r = 0$ of Minkowski spacetime.

An important question, then, is whether such singularities develop even when a general model is considered, and if so under what conditions. To answer this, it is first necessary to characterize precisely what one means by a spacetime singularity for a general spacetime. Then it is seen that singularities must exist for a very wide class of spacetimes under a reasonable general set of conditions. Such singularities may develop as the end state of the gravitational collapse of a massive star, or in cosmological situations such as the origin of the universe.

The first point to note here is by very definition, the metric tensor must be well-defined at all regular points of the spacetime. This is no longer true at a spacetime singularity such as those discussed above and a singularity cannot be regarded as a regular point of spacetime but is a boundary point attached to M. This causes difficulties when one tries to characterize a singularity by the criterion that the curvatures must blow up near the singularity. The trouble is, since the singularity is not part of the spacetime, it is not possible to define its neighborhood in the usual sense to discuss the behavior of curvature quantities in that region. One may try to characterize a singularity in terms of

the divergence of components of the Riemann curvature tensor along nonspace-like trajectories. Then the trouble is that the behavior of such components will, in general, change with the frames used, so this is not of much help. One can try the curvature scalars or the scalar polynomials in the metric and the Riemann tensor and require them to achieve unboundedly large values. This is encountered in Schwarzschild and Friedmann models. However, it is possible that such a divergence occurs only at infinity for a given nonspace-like curve. In general, it looks reasonable to demand that some sort of curvature divergence must take place along the nonspace-like curves which encounter a singularity. However, a general characterization of singularity in terms of the curvature divergence runs into various difficulties.

Considering these and similar situations, the occurrence of nonspace-like geodesic incompleteness is generally agreed upon as the criterion for the existence of a singularity for a spacetime. This may not cover all types of singular behaviors possible, but it is clear that if a spacetime manifold contains incomplete nonspace-like geodesics, there is a definite singular behavior present, as a time-like observer or a photon suddenly disappears from the spacetime after a finite amount of proper time or after a finite value of the affine parameter. The singularity theorems which result from an analysis of gravitational focusing and global properties of a spacetime prove this incompleteness property for a wide class of spacetimes under a set of rather general conditions.

Physically, a singularity in any physics theory typically means that the theory breaks down either in the vicinity or at the singularity. This means that a broader and more comprehensive theory is needed, demanding a revision of the given theory. Similar reasoning apply to spacetime singularities, which may be taken to imply that a quantum gravity description is warranted in those regions of the universe, rather than using merely a classical framework.

20.3 Gravitational Focusing

Simple model solutions such as Schwarzschild and FRW universes give very useful indications as to what is possible in general relativity, as opposed to Newtonian gravity. In particular, these solutions are important indicators on the existence and nature of the spacetime singularities.

The key to occurrence of singularities in these solutions is really the gravitational focusing that the matter fields and spacetime curvature cause in the congruences of null and time-like curves, which represent the light paths and the material particle trajectories in any given spacetime universe. It would then be important to know how general and generic such a feature and property is for a general spacetime.

The matter fields with positive energy density which create the curvature in spacetime affect the causality relations in a spacetime and create focusing in families of nonspace-like trajectories. The phenomenon that occurs here is that matter focuses the nonspace-like geodesics of the spacetime into pairs of *focal points* or *conjugate points*. The property of conjugate points is that if p, q are two conjugate points along a nonspace-like geodesic, then p and q must be time-like related. Now, there are three-dimensional null hypersurfaces such as the boundary of the future of an event such that no two points of such a hypersurface can be joined by a time-like curve. Thus, the null geodesic generators of such surfaces cannot contain conjugate points and they must leave the hypersurface before encountering any conjugate point. This puts strong limits on such null surfaces and the singularity theorems result from an analysis of such limits.

If we consider a congruence of time-like geodesics in the spacetime, this is a family of curves and through each event there passes precisely one time-like geodesic trajectory. Choosing the curves to be smooth, this defines a smooth time-like vector field on the spacetime. The rate of change of volume expansion for a given congruence of time-like geodesics can be written as

$$\frac{\mathrm{d}\theta}{\mathrm{d}\tau} = -R_{lk}V^l V^k - \tfrac{1}{3}\theta^2 - 2\sigma^2 + 2\omega^2 ,$$

where, for a given congruence of time-like (or null) geodesics, the quantities θ, σ, and ω are *expansion*, *shear*, and *rotation* tensors, respectively. The above equation is called the *Raychaudhuri equation* [20.5], which describes the rate of change of the volume expansion as one moves along the time-like geodesic curves in the congruence.

We note that the second and third term on the right-hand side involving θ and σ are always positive. For the term $R_{ij}V^iV^j$, by Einstein equations this can be written as

$$R_{ij}V^iV^j = 8\pi\left[T_{ij}V^iV^j + \tfrac{1}{2}T\right].$$

The term $T_{ij}V^iV^j$ above represents the energy density measured by a time-like observer with unit tangent V^i, which is the four-velocity of the observer. For all reasonable classical physical fields this energy density is generally taken as nonnegative and we may assume for all time-like vectors V^i

$$T_{ij}V^iV^j \geq 0\,.$$

Such an assumption is called the *weak energy condition.* It is also considered reasonable that the matter stresses will not be so large as to make the right-hand side of the equation above negative. This will be satisfied when the following is satisfied: $T_{ij}V^iV^j \geq -\frac{1}{2}T$. Such an assumption is called the *strong energy condition* and it implies that for all time-like vectors V^i, $R_{ij}V^iV^j \geq 0$. By continuity it can be argued that the same will then hold for all null vectors as well. Both the strong and weak energy conditions will be valid for well-known forms of matter such as the perfect fluid provided that the energy density ρ is nonnegative and there are no large negative pressures which are bigger or comparable to ρ.

With the strong energy condition, the Raychaudhuri equation implies that the effect of matter on spacetime curvature causes a focusing effect in the congruence of time-like geodesics due to gravitational attraction. This causes neighboring geodesics in the congruence to cross each other to give rise to caustics or conjugate points. Such a separation between nearby time-like geodesics is governed by the geodesic deviation equation

$$D^2Z^j = -R^j{}_{kil}V^kZ^iV^l\,,$$

where Z^i is the separation vector between nearby geodesics of the congruence. Solutions of the above equation are called the *Jacobi fields* along a given time-like geodesic.

Suppose now that γ is a time-like geodesic, then two points p and q along γ are called *conjugate points* if there exists a Jacobi field along γ which is not identically zero but vanishes at p and q. From the Raychaudhuri equation given above it is clear that the occurrence of conjugate points along a time-like geodesic is closely related to the behavior of the expansion parameter θ of the congruence. In fact, it can be shown that the necessary and sufficient condition for a point q to be conjugate to p is that for the congruence of time-like geodesics emerging from p, we must have $\theta \to -\infty$ at q (see, for example, [20.1]). The conjugate points along null geodesics are also similarly defined.

20.4 Geodesic Incompleteness

It was widely believed that for more general solutions of the Einstein equations which incorporate several other physical features and are not necessarily symmetric, the existence of singularities would be avoided (see, e.g., [20.6]; and for recent development and reviews we refer to [20.7] and [20.8]). Further investigations, however, showed that singularities in the form of geodesic incompleteness do exist for general spacetimes. These results used the gravitational focusing considerations mentioned above and global properties of a general spacetime.

The behavior of the expansion parameter θ is governed by the Raychaudhuri equation as pointed out above. Consider, for example, the situation when the spacetime satisfies the strong energy condition and the congruence of time-like geodesics is hypersurface orthogonal. Then $\omega_{ij} = 0$ and the corresponding term ω^2 vanishes. Further, the expression for the covariant derivative of ω_{ij} implies that it must vanish for all future times as well. It follows that we must then have $\mathrm{d}\theta/\mathrm{d}\tau \leq -(\theta^2/3)$, which means that the volume expansion parameter must be necessarily decreasing along the time-like geodesics. If θ_0 denotes the initial expansion then the above can be integrated as $\theta^{-1} \geq \theta_0^{-1} + \tau/3$. It is clear from this that if the congruence is initially converging and θ_0 is negative, then $\theta \to -\infty$ within a proper time distance $\tau \leq 3/\mid\theta_0\mid$.

It then follows that if M is a spacetime satisfying the strong energy condition and S is a space-like hypersurface with $\theta < 0$ at $p \in S$, then if γ is a time-like geodesic of the congruence orthogonal to S passing through p there exists a point q conjugate to S along γ within a proper time distance $\tau \leq 3/\mid\theta\mid$. This is provided that γ can be extended to that value of the proper time.

The basic implication of the above results is that once a convergence occurs in a congruence of time-like geodesics, the conjugate points or the caustics must develop in the spacetime. These can be interpreted as the singularities of the congruence. Such singularities could occur even in Minkowski spacetime and similar other perfectly regular spacetimes. However, when combined with certain causal structure properties of spacetime, the results above imply the

existence of singularities in the form of geodesic incompleteness. One could similarly discuss the gravitational focusing effect for the congruence of null geodesics or for null geodesics orthogonal to a space-like two-surface.

There are several singularity theorems available that establish nonspace-like geodesic incompleteness for a spacetime under different sets of conditions and are applicable to different physical situations. However, the most general of these is the Hawking–Penrose theorem [20.9], which is applicable in both the collapse situation and cosmological scenario. The main idea of the proof of such a theorem is the following. Using the causal structure analysis it is shown that there must be maximal length time-like curves between certain pairs of events in the spacetime. Now, a causal geodesic which is both future and past complete must contain pairs of conjugate points if M satisfies an energy condition. This is then used to draw the necessary contradiction to show that M must be nonspace-like geodesically incomplete.

Theorem 20.1

A spacetime (M, g) cannot be time-like and null geodesically complete if the following are satisfied:

1. $R_{ij}K^iK^j \geq 0$ for all nonspace-like vectors K^i.
2. The generic condition is satisfied, that is, every nonspace-like geodesic contains a point at which $K_{[i}R_{j]el[m}K_{n]}K^eK^l \neq 0$, where K is the tangent to the nonspace-like geodesic.
3. The chronology condition holds on M; that is, there are no closed time-like curves in the spacetime.
4. There exists in M either a compact achronal set (i. e., a set no two points of which are time-like related) without edge or a closed trapped surface, or a point p such that for all past directed null geodesics from p, eventually θ must be negative.

The main idea of the proof is the following. One shows that the following cannot hold simultaneously:

a) Every inextendible nonspace-like geodesic contains pairs of conjugate points.
b) The chronology condition holds on M.
c) There exists an achronal set S in M such that $E^+(S)$ or $E^-(S)$ is compact.

In the above, E^+ and E^- indicate the future and past horismos for the set S (for further definitions and details we refer to *Hawking* and *Ellis* [20.1], or *Joshi* [20.10]).

We note that while geodesic incompleteness, as a definition of spacetime singularities, allows various theorems to be proved on the existence of singularities, it does not capture all possible singular behaviors for a spacetime. It also does not imply that the singularity predicted is necessarily a physically relevant powerful curvature singularity. It does, of course, include many cases where that will be the case. Below, we discuss such a scenario and the criterion for the singularity to be physically relevant and important.

20.5 Strong Curvature Singularities

As we saw above, the existence of an incomplete nonspace-like geodesic or the existence of an inextendible nonspace-like curve which has a finite length as measured by a generalized affine parameter, implies the existence of a spacetime singularity. The *generalized affine length* for such a curve is defined as [20.1]

$$L(\lambda) = \int_0^a \left[\sum_{i=0}^{3}(X^i)^2\right]^{1/2} \mathrm{d}s\,,$$

which is a finite quantity. The X^is are the components of the tangent to the curve in a parallel propagated tetrad frame along the curve. Each such incomplete curve defines a boundary point of the spacetime which is a singularity.

The important point now is, in order to call this a genuine physical singularity, one would typically like to associate such a singularity with unboundedly growing spacetime curvatures. If all the curvature components and the scalar polynomials formed out of the metric and the Riemann curvature tensor remained finite and well-behaved in the limit of approach to the singularity along an incomplete nonspace-like curve, it may be possible to remove such a singularity by extending the spacetime when the differentiability requirements are lowered [20.11].

There are several ways in which such a requirement can be formalized. For example, a *parallely propagated curvature singularity* is the one which is the end point of at least one nonspace-like curve on which the com-

ponents of the Riemann curvature tensor are unbounded in a parallely propagated frame. On the other hand, a *scalar polynomial singularity* is the one for which a scalar polynomial in the metric and the Riemann tensor takes an unboundedly large value along at least one nonspace-like curve which has a singular end point. This includes cases such as Schwarzschild singularity, where the Kretschmann scalar $R^{ijkl}R_{ijkl}$ blows up in the limit as $r \to 0$.

What is the guarantee that such curvature singularities will at all occur in general relativity? The answer to this question for the case of parallely propagated curvature singularities is provided by a theorem of *Clarke* [20.12], which establishes that for a globally hyperbolic spacetime M which is C^{0-} inextendible, when the Riemann tensor is not very specialized in the sense of not being type-D and electrovac at the singular end point, then the singularity must be a parallely propagated curvature singularity.

Curvature singularities to be characterized below, also arise for a wide range of spacetimes involving gravitational collapse. This physically relevant class of singularities, called the *strong curvature singularities* was defined and analyzed by *Tipler* [20.13], *Tipler* et al. [20.14], and *Clarke* and *Królak* [20.15]. The idea here is to define a physically all embracing strong curvature singularity in such a way so that all the objects falling within the singularity are destroyed and crushed to zero volume by the infinite gravitational tidal forces. The extension of spacetime becomes meaningless for such a strong singularity which destroys to zero size all the objects terminating at the singularity. From this point of view, the strength of singularity may be considered crucial to the issue of classically extending the spacetime, thus avoiding the singularity. This is because for a strong curvature singularity defined in the above sense, no continuous extension of the spacetime may be possible.

20.6 Can We Avoid Spacetime Singularities?

Given the scenario above, it is now clear that spacetime singularities are an inevitable feature for most of the physically reasonable models of universe and gravitational systems within the framework of the Einstein theory of gravity. It is also seen that near such a spacetime singularity, the classical description that predicted it must itself break down. The existence of singularities in most of the classical theories of gravity, under reasonable physical conditions, imply that in a sense the Einstein gravity itself predicts its own limitations, namely that it predicts regions of the universe where it must breakdown and a new and revised physical description must take over.

As the curvatures and all other physical quantities must diverge near such a singularity, the quantum effects associated with gravity are very likely to dominate such a regime. It is possible that these may resolve the classical singularity itself. However, we currently do not have any viable and consistent quantum theory of gravity despite serious attempts. Therefore, the issue of resolution of singularities as produced by classical gravity remains open.

The other possibility is, of course, that some of the assumptions of the singularity theorems may be violated so as to avoid the singularity occurrence. Even when these are fairly general, one could inquire whether some of them could actually breakdown and do not hold in physically realistic models. This could save us from the occurrence of singularities at the classical level itself. Such possibilities mean a possible violation of causality in the spacetime, or no trapped surfaces occurring in the dynamical evolution of the universe, or possible violation of energy conditions.

The singularity theorem stated above and also other singularity theorems contain the assumption of causality or strong causality, or some other suitable causality condition. Then the alternative is that causality may be violated rather than a singularity occurring in the spacetime. So the implication of the singularity theorem stated above is that when there is enough matter present in the universe, either the causality is violated or a boundary point must exist for the spacetime. In the cosmological case, such stress-energy density will be provided by microwave background radiation, or in the case of stellar collapse trapped surfaces may form [20.16], providing a condition leading to the formation of a singularity.

The Einstein equations by themselves do not rule out causality violating configurations which really depend on the global topology of the spacetime. Hence the question of causality violations versus spacetime singularity needs a careful examination as to whether

causality violation could offer an alternative to singularity formation. Similarly, it must also be inquired whether the violation of energy conditions or nonoccurrence of trapped surfaces may be realized so as to achieve singularity avoidance in a spacetime. We briefly discuss some of these points below.

20.7 Causality Violations

The causal structure in a spacetime specifies what events can be related to each other by means of time-like or light signals. A typical causality violation would mean that an event could be in its own past, which is contrary to our normal understanding of time, and that of past and future. This has been examined in considerable detail in general relativity, and that no causality violation takes place in the spacetime is one of the important assumptions used by singularity theorems. However, general relativity allows for situations where causality violation is permitted in a spacetime. The Gödel solution [20.17] allows the existence of a closed time-like curve through every point of the spacetime.

One would, of course, like to rule out if possible causality violations on physical grounds, treating them as very pathological behavior in that in such a case one would be able to enter one's own past. However, as they are allowed in principle in general relativity, so one must rule them out only by an additional assumption. The question then is, can one avoid spacetime singularities if one allows for the violation of causality? This has been considered by researchers and it was seen that the causality violation in its own right creates spacetime singularities again under certain reasonable conditions. Thus, this path of avoiding spacetime singularities does not appear to be very promising.

Specifically, the question of finite causality violations in asymptotically flat spacetime was examined by *Tipler* [20.13, 18, 19]. This showed that the causality violation in the form of closed time-like lines is necessarily accompanied by incomplete null geodesics, provided the strong energy condition is satisfied for all null vectors and the generic condition is satisfied. It was assumed that the energy density ρ has a positive minimum along past directed null geodesics.

There is, in fact, a hierarchy of causality conditions available for a spacetime. It may be causal in the sense of having no closed nonspace-like curves. However, given an event, future directed nonspace-like curves from the same event could return to its arbitrarily close neighborhood in the spacetime. This is as bad as a causality violation itself. The higher-order causality conditions such as strong causality and stable causality rule out such behavior. Of the higher-order causality conditions, much physical importance is attached to stable causality, which ensures that if M is causal, its causality should not be disturbed with small perturbations in the metric tensor. Presumably, general relativity is a classical approximation to some, as yet unknown, quantum theory of gravity in which the value of the metric at a point will not be exactly known and small fluctuations in the value must be taken into account.

Results on causality violations and higher-order causality violations with reference to occurrence of singularities were obtained by *Joshi* [20.20], and *Joshi* and *Saraykar* [20.21], who showed that the causality violations must be accompanied by singularities even when the spacetime is causal but the higher-order causality conditions are violated. Thus we know that for a causal spacetime, the violations of higher-order causality conditions give rise to spacetime singularities. Another question examined was that of the measurement of causality violating sets when such a violation occurs. It turns out that in many cases, the causality violating sets in a spacetime will have a zero measure, and thus such a causality violation may not be taken very seriously. Also, *Clarke* and *Joshi* [20.22] studied global causality violation for a reflecting spacetime and the theorems of *Kriele* [20.23] improved some of the conditions under which the results on chronology violations implying the singularities have been obtained. Also, global causality violating spacetimes were studied by *Clarke* and *de Felice* [20.24]. What we discussed above implies that if the causality of M breaks down with the slightest perturbation of the metric, then this must be accompanied by the occurrence of spacetime singularities.

As a whole, the above results imply that violating either causality or any of the higher-order causality conditions may not be considered a good alternative to the occurrence of spacetime singularities. There are also philosophical problems connected with the issue of causality violation, such as entering one's own past. However, even if one allowed for the causality violations, the above results show that these are necessarily accompanied by spacetime singularities again.

20.8 Energy Conditions and Trapped Surfaces

Another possibility to avoid singularities is a possible violation of energy conditions. This is another of the assumptions in the proofs for singularity theorems. In fact, this possibility has also been explored in some detail and it turns out that as long as there is no gross or very powerful violation of energy conditions over global regions in the universe, this would not help avoid singularities either.

For example, the energy condition could be violated locally at certain spacetime points, or in certain regions of spacetimes due to peculiar physics there. However, as long as it holds on an average, in the sense that the stress-energy density is positive in an integrated sense, then spacetime singularities still occur (for a discussion and references, see, e.g., [20.10]).

On a global scale, there is evidence now that the universe may be dominated by a dark energy field. There is no clarity as to what exactly such a field would be and what would be its origin. It could be due to scalar fields or ghost fields floating in the universe, or due to a nonzero positive cosmological constant present in the Einstein equations. In such a case, the weak or strong energy conditions may be violated depending on the nature of these exotic fields. However, in the earlier universe of a matter dominated phase, the positive matter fields would again dominate, thus respecting the energy conditions even if they are violated at the present epoch.

Again, the above discussion is in the context of a cosmological scenario. When it comes to the gravitational collapse of massive stars, clearly their density and overall energy content are dominated by the ordinary matter fields with which we are much more familiar. Such matter certainly respects the energy conditions modulo with some minor violations if at all any. Thus for gravitational collapse of massive stars, one would expect the energy conditions to hold and the conclusions on singularity occurrence stated above would apply.

Yet another possibility to avoid singularity is to avoid trapped surfaces occurring in the spacetime. Indeed, such a route can give rise to geodesically complete spacetimes, as was shown by *Senovilla* [20.25]. As for the cosmological scenario, basically this means and amounts to the condition that the matter energy densities must fall off sufficiently rapidly on any given space-like surface, and in an averaged sense, in order to avoid the cosmological trapped surfaces. Whether such a condition is realizable in the universe would have to be checked through observational tests. A sufficiently uniform energy density, such as, say, the microwave background radiation could in turn cause cosmic trapping. As for massive stars, the densities are, of course, very high indeed and would only grow, for example, in a gravitational collapse. Therefore, in collapse scenarios, the trapped surfaces are unlikely to be avoided.

Further to the above considerations, if we accept on the whole that spacetime singularities do occur under fairly general conditions in the framework of the Einstein theory of gravity, or for classical gravity in general, then we must consider physical implications and consequences of such a scenario for physics and cosmology. As we noted earlier, two main arenas of physical relevance where spacetime singularities will be of interest are the cosmological situation and the gravitational collapse scenarios.

20.9 Fundamental Implications and Challenges

The existence of spacetime singularities in Einstein gravity and other classical theories of gravitation poses intriguing challenges and fundamental questions in physics as well as cosmology. These would have far-reaching consequences for our current understanding of the universe and how we try to model it further, as we shall try to bring out in rest of this article.

The inevitable existence of singularities for wide classes of rather general models of spacetimes means that the classical gravity evolutions necessarily give rise to regions in the spacetime universe where the densities and spacetime curvatures would really grow arbitrarily high without any bounds, and where all other relevant physical parameters would also diverge.

To take the first physical scenario, such a phenomenon in cosmology would correspond to a singularity that will represent the origin of the universe. Secondly, whenever locally a large quantity of matter and energy collapses under the force of its own gravity, a singularity will occur. This later situation will be effectively realized in the gravitational collapse of massive stars in the universe, which collapse and shrink

catastrophically under their self-gravity, when the star has exhausted its nuclear fuel within that earlier supplied the internal pressures to halt the infall due to gravity.

Over the past decades, once the existence of spacetime singularities was accepted, there have been major efforts to understand the physics in the vicinity of the same. In the cosmological case, this has given rise to an entire physics of the early universe, which depicts the few initial moments immediately after the big bang singularity from which the universe is supposed to have emerged. The complexities in this case have been enormous, both physics-wise, and conceptually. The physics complexities arise because when trying to understand physics close to the hot big bang singularity, we are dealing with the highest energy scales, never seen earlier in any laboratory physics experiments. Our particle physics theories are then to be stretched to the extreme where there is no definite or unique framework available to deal with these phenomena. Understanding early universe physics has, of course, very big consequences in that it governs the most important physical phenomena such as the later galaxy formation in the universe and other issues related to the large-scale structure of the universe.

As for the conceptual issues, simultaneous big bang singularity gives rise to a host of problems and puzzles. One of these is the *horizon problem* which arises due to the causal structure of this spacetime. Distinct regions of the universe simply cannot interact with each other due to the cosmic horizons and it becomes extremely difficult to explain the average overall current homogeneity of the universe on a large enough scale. There are also other issues such as why the current universe looks so flat, which is called the *flatness* problem. As a possible means to resolve these dilemmas inflationary models for the early universe have been proposed, various facets of which are still very much under an active debate.

The key issue, as far as big bang singularity is concerned, is that it happened only once in the past and there is no way to probe it any further other than current observations on the universe and their extrapolation in the past. One must look deeper and deeper into space and back into time to understand the nature and physics of this early universe singularity.

As we mentioned above, the other class of such spacetime singularities will occur in the gravitational collapse of massive stars in the universe. Unlike the big bang, such a singularity will occur whenever a massive star in the universe collapses. This is, therefore, more amenable to observational tests.

There are rather fundamental cosmic conundrums associated with singularities of gravitational collapse. One of the most intriguing of these is the question of whether such a singularity will be visible to external faraway observers in the universe. The big bang singularity is visible to us in principle, as we get to see the light from the same. However, as we discuss below, the singularity of collapse can sometimes be hidden below the event horizons of gravity, and are therefore not visible. The possibility that all singularities of collapse will be necessarily hidden inside horizons is called the *cosmic censorship conjecture*. As we discuss below this is not yet proved, and, in fact, singularities of collapse can be visible under many physical circumstances.

When visible or naked singularities develop in gravitational collapse, they give rise again to extremely intriguing physical possibilities and problems. The opportunity offered in that case is that we may have the possibility to observe the possible ultra-high energy physical processes occurring in such a region of the universe, including quantum gravity effects. Such observations of ultra-high energy events in the universe could provide observational tests and guide our efforts for a possible quantum theory of gravity. However, a conundrum that is presented is whether this would break the so-called *classical predictability* of the spacetime universe. We shall discuss this further below. On the other hand, even when the singularity is necessarily hidden within a black hole, this still gives rise to profound puzzles such as the *information paradox*, issues with *unitarity* and other such problems. So the point is that, even if the cosmic censorship was correct and all singularities were hidden inside black holes only, we shall still be faced with many deep paradoxes, which are not unique to naked singularities only.

It would be only reasonable to say that all these deep physical as well as conceptual issues are closely connected with the existence and formation of spacetime singularities in the dynamical gravitational processes taking place in the universe. While the big bang singularity happened only once in the past, the singularities of collapse have, in fact, a repeated occurrence, and hence they possess an interesting observational perspective and potential. We shall, therefore, discuss the same in some detail below, while also providing the key ingredients of black hole physics in the process.

20.10 Gravitational Collapse

When a massive star of more than about ten solar masses exhausts its internal nuclear fuel, it is believed to enter the stage of continual gravitational collapse without any final equilibrium state. The star then goes on shrinking in its radius, reaching higher and higher densities. What would be the final fate of such an object according to the general theory of relativity? This is one of the central questions in relativistic astrophysics and gravitation theory today. It is suggested that the ultra-dense object that forms as a result of the collapse could be a black hole in space and time from which not even light rays escape. Alternatively, if an event horizon of gravity fails to cover the final super ultra-dense crunch, it could be a visible singularity in the spacetime which could causally interact with the outside universe and from which region the emissions of light and matter may be possible.

The issue is of importance from both the theoretical as well as the observational point of view. At the theoretical level, working out the final fate of collapse in general relativity is crucial to the problem of asymptotic predictability, namely, whether the singularities forming as the collapse end state will be necessarily covered by the event horizons of gravity. Such a *censorship hypothesis* remains fundamental to the theoretical foundations of black hole physics and its many recent astrophysical applications. These include the area theorem for black holes, laws of black hole thermodynamics, Hawking radiation, predictability in a spacetime, and on the observational side the accretion of matter by black holes, massive black holes at the center of galaxies, etc. On the other hand, the existence of visible or naked singularities offers a new approach to these issues requiring modification and reformulation of our usual theoretical conception of black holes.

We mention and discuss below some of the recent developments in these directions, examining the possible final fate of gravitational collapse. To investigate this issue, dynamical collapse scenarios have been examined in the past decade or so for many cases such as clouds composed of dust, radiation, perfect fluids, or matter with more general equations of state (see, e.g., [20.26] for references and details). We discuss these developments and the implications for a possible formulation of cosmic censorship are indicated, mentioning the open problems in the field.

20.11 Spherical Collapse and the Black Hole

To understand the final fate of a massive gravitationally collapsing object we first outline here the spherically symmetric collapse situation. Although this is an idealization, the advantage is that one can solve the case analytically to obtain exact results when matter is homogeneous dust. In fact, the basic motivations for the idea and theory of black holes come from this collapse model, first worked out by *Oppenheimer* and *Snyder* [20.27] and *Datt* [20.28].

Consider a gravitationally collapsing spherical massive star. The interior solution for the object depends on the properties of matter, equations of state, and the physical processes taking place within the stellar interior. However, assuming the matter to be pressureless dust allows us to solve the problem analytically, which provides important insights. The energy–momentum tensor is given by $T^{ij} = \rho u^i u^j$ and the Einstein equations are to be solved for the spherically symmetric metric. The metric potentials can be solved and the interior geometry of the collapsing dust ball is given by

$$ds^2 = -dt^2 + R^2(t)\left[\frac{dr^2}{1-r^2} + r^2 d\Omega^2\right],$$

where $d\Omega^2 = d\theta^2 + \sin^2\theta\, d\phi^2$ is the metric on the two-sphere. The interior geometry of the cloud matches at the boundary $r = r_b$ with the exterior Schwarzschild spacetime.

The basic features of such a configuration are given in Fig. 20.1. The collapse is initiated when the star surface is outside the Schwarzschild radius $r = 2m$ and a light ray from the surface of the star can escape to infinity. However, once the star has collapsed below $r = 2m$, a black hole region of no escape develops in the spacetime which is bound by the event horizon at $r = 2m$. Any point in this empty region represents a *trapped surface*, a two-sphere for which both the outgoing and ingoing families of null geodesics emitted from this point converge and thus no light comes out

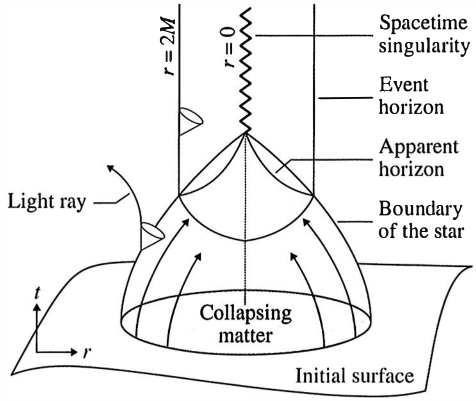


Fig. 20.1 The gravitational collapse of a spherically symmetric homogeneous dust cloud. The event horizon forms prior to the singularity and the collapse end state is a black hole

of this region. Then, the collapse to an infinite density and curvature singularity at $r = 0$ becomes inevitable in a finite proper time as measured by an observer on the surface of the star. The black hole region in the resulting vacuum Schwarzschild geometry is given by $0 < r < 2m$ with the event horizon as the outer boundary at which the radial outwards photons stay where they are but all the others are dragged inwards. No information from the black hole can propagate outside $r = 2m$ to observers far away. We thus see that the collapse gives rise to a black hole in the spacetime which covers the resulting spacetime singularity. The ultimate fate of the star undergoing such a collapse is an infinite curvature singularity at $r = 0$, completely hidden within the black hole. No emissions or light rays from the singularity go out to an observer at infinity and the singularity is causally disconnected from the outside spacetime.

The question now is whether one could generalize these conclusions on the occurrence of spacetime singularity in collapse for more general forms of matter or for nonspherical situations, or possibly for small perturbations away from spherical symmetry. It is known using the stability of Cauchy development in general relativity that the formation of trapped surfaces is indeed a stable property when departures from spherical symmetry are taken into account. The argument essentially is the following. Considering a spherical collapse evolution from given initial data on a partial Cauchy surface S, we find the formation of trapped surfaces T in the form of all the spheres with $r < 2m$ in the exterior Schwarzschild geometry. The stability of Cauchy development then implies that for all initial data sufficiently near the original data in the compact region $J^+(S) \cap J^-(T)$, where J^+ and J^- denote the causal futures or pasts of S, respectively, the trapped surfaces still must occur. Then, the curvature singularity of spherical collapse also turns out to be a stable feature, as implied by the singularity theorems which show that the closed trapped surfaces always imply the existence of a spacetime singularity under reasonable relatively general conditions.

20.12 Cosmic Censorship Hypothesis

Real stars in the universe are not made of pressureless homogeneous dust. They are inhomogeneous, typically with density higher at the center, may have nontrivial matter with an equation of state that is as yet unknown, and there is spin. Will a physically realistic collapse of such a star necessarily end up in the black hole final state only, just as in the idealized case of the Oppenheimer–Snyder–Datt model above? In other words, while more general gravitational collapse will also end up in a spacetime singularity, the question is whether the singularity will be again necessarily covered inside an event horizon of gravity.

In fact, there is no proof available that such a singularity will continue to be hidden within a black hole and remain causally disconnected from outside observers, even when the collapse is not exactly spherical or when the matter does not have the form of exact homogeneous dust. Therefore, in order to generalize the notion of black holes to more general gravitational collapse situations, it becomes necessary to rule out such naked or visible singularities by means of an explicit assumption. This is stated as the *cosmic censorship hypothesis* [20.29], which essentially says that if S is a partial Cauchy surface from which the collapse commences, then there are no naked singularities to the future of S which could be seen from the future null infinity. This is true for spherical homogeneous dust collapse, where the resulting spacetime is future asymptotically predictable and the censorship holds. In such a case, the breakdown of physical theory at the space-

time singularity does not disturb prediction in future for the outside asymptotically flat region.

The corresponding scenario for other more general collapse situations, when inhomogeneities or nonsphericity and such other physically realistic features are allowed for must be investigated. The answer in general is not known either as a proof of the future asymptotic predictability for general spacetimes or in the form of any general theorem on cosmic censorship. It is clear that the assumption of censorship in a suitable form is crucial to the basic results in black hole physics. Actually when one considers the gravitational collapse in a generic situation, the very existence of black holes requires this hypothesis.

To establish censorship by means of a rigorous proof certainly requires a much more precise formulation of the hypothesis. The statement that the result of complete gravitational collapse must be a black hole only and not a naked singularity, or all singularities of collapse must be hidden inside black holes, is not rigorous enough. Because under general circumstances, the censorship or asymptotic predictability is false as one could always choose a spacetime manifold with a naked singularity which would be a solution to Einstein's equations if we define $T_{ij} \equiv (1/8\pi)G_{ij}$. So certain conditions on the stress–energy tensor are required at the minimum, say, for example, an energy condition. However, to obtain an exact set of conditions on matter fields to prove the censorship hypothesis turns out to be an extremely difficult task that has as yet not been accomplished.

The requirements in black hole physics and general predictability have led to several different possible formulations of cosmic censorship hypothesis, none of which has been proved as yet. *Weak cosmic censorship*, or asymptotic predictability, postulates that the singularities of gravitational collapse cannot influence events near the future null infinity. The other version called the *strong cosmic censorship* is a general predictability requirement on any spacetime, stating that all physically reasonable spacetimes must be globally hyperbolic (see, e.g., [20.30]). *Global hyperbolicity* here means that we must be able to predict the entire future and past evolutions in the universe by means of the Einstein equations, given the initial data on a three-dimensional space-like hypersurface in spacetime.

However, on further analysis it becomes clear that such formulations need much more sharpening if any concrete proof is to be obtained at all. In fact, as for the cosmic censorship, it is a major problem in itself to find a satisfactory and mathematically rigorous formulation of what it is physically desired to achieve. Presently, there is no general proof available for any suitably formulated version of weak censorship. The main difficulty seems to be that the event horizon is a feature depending on the whole future behavior of the solution over an infinite time period, but the present theory of quasi-linear hyperbolic equations guarantees the existence and regularity of solutions over only a finite time internal. It is clear that even if true, any proof for a suitable version of weak censorship seems to require much more knowledge on general global properties of Einstein equations than is currently known.

To summarize the situation, cosmic censorship is clearly a crucial assumption underlying all of black hole physics and gravitational collapse theory, and related important areas. The first major task here would be to formulate rigorously a satisfactory statement for cosmic censorship, which if not true would throw black hole dynamics into serious doubt. This is why censorship is one of the most important open problems for gravitation theory today. No proof, however, seems possible unless some major theoretical advances by way of mathematical techniques and understanding the global structure of Einstein equations are made, and the direction needed for such theoretical advances is far from clear at present.

We, therefore, conclude that the first and foremost task at the moment is to carry out a detailed and careful examination of various gravitational collapse scenarios to examine for their end states. It is only such an investigation of more general collapse situations which could indicate what theoretical advances to expect for any proof, and what features to avoid while formulating the cosmic censorship. Basically, we still do not have sufficient data and information available on the various possibilities for gravitationally collapsing configurations so as to decide one way or other on the issue of censorship.

In recent years, many investigations have been carried out from such a perspective on gravitational collapse, either for inhomogeneous dust collapse or with more general matter fields. It turns out that the collapse outcome is not always a black hole and the naked singularity final state can arise in a variety of situations. In the next sections we discuss some of these developments.

20.13 Inhomogeneous Dust Collapse

Since we are interested in collapse, we require that the spacetime contains a regular initial space-like hypersurface on which the matter fields as represented by the stress-energy tensor T_{ij} have compact support and all physical quantities are well-behaved on this surface. Also, the matter should satisfy a suitable energy condition and the Einstein equations. We say that the spacetime contains a naked singularity if there is a future directed nonspace-like curve which reaches a far away observer at infinity in future, which in the past terminates at the singularity.

As an immediate generalization of the Oppenheimer–Snyder–Datt homogeneous dust collapse, one could consider the collapse of inhomogeneous dust and examine the nature and structure of resulting singularity with special reference to censorship and the occurrence of black holes and naked singularities. The main motivation to discuss this situation is that it provides a clear picture in an explicit manner of what is possible in gravitational collapse. One could ask how the conclusions given above for homogeneous collapse are modified when the inhomogeneities of matter distribution are taken into account. Clearly, it is important to include effects of inhomogeneities because typically a physically realistic collapse would start from an inhomogeneous initial data with a centrally peaked density profile.

This question of inhomogeneous dust collapse has attracted the attention of many researchers and it is seen that the introduction of inhomogeneities leads to a rather different picture of gravitational collapse. It turns out that while homogeneous collapse leads to black hole formation, the introduction of any physically realistic inhomogeneity, e.g., the density peaked at the center of the cloud and slowly decreasing away, leads to a naked singularity final state for the collapse. This is certainly an intriguing result implying that the black hole formation in gravitational collapse may not be such a stable phenomenon as was thought to be the case.

The problem was investigated in detail using the Tolman–Bondi–Lemaître models, which describe gravitational collapse of an inhomogeneous spherically symmetric dust cloud [20.31]. This is an infinite dimensional family of asymptotically flat solutions of Einstein equations, which is matched to the Schwarzschild spacetime outside the boundary of the collapsing star. The Oppenheimer–Snyder–Datt model is a special case of this class of solutions.

It is seen that the introduction of inhomogeneities leads to a rather different picture of gravitational collapse. The metric for spherically symmetric collapse of inhomogeneous dust, in comoving coordinates (t, r, θ, ϕ) is given by

$$ds^2 = -dt^2 + \frac{R'^2}{1+f} dr^2 + R^2 (d\theta^2 + \sin^2\theta \, d\phi^2)$$

$$T^{ij} = \epsilon \delta^i_t \delta^j_t , \quad \epsilon = \epsilon(t, r) = \frac{F'}{R^2 R'} ,$$

where T^{ij} is the stress–energy tensor, ϵ is the energy density, and R is a function of both t and r given by

$$\dot{R}^2 = \frac{F}{R} + f .$$

Here the dot and prime denote partial derivatives with respect to the parameters t and r, respectively. As we are considering collapse, we require $\dot{R}(t, r) < 0$. The quantities F and f are arbitrary functions of r and $4\pi R^2(t, r)$ is the proper area of the mass shells. The physical area of such a shell at $r = \text{const.}$ goes to zero when $R(t, r) = 0$. For the gravitational collapse situation, we take ϵ to have compact support on an initial space-like hypersurface and the spacetime is matched at some $r = \text{const.} = r_c$ to the exterior Schwarzschild field with the total Schwarzschild mass $m(r_c) = M$ enclosed within the dust ball of coordinate radius of $r = r_c$. The apparent horizon in the interior dust ball lies at $R = F(r)$.

Using this framework, the nature of the singularity $R = 0$ can be examined. In particular, the problem

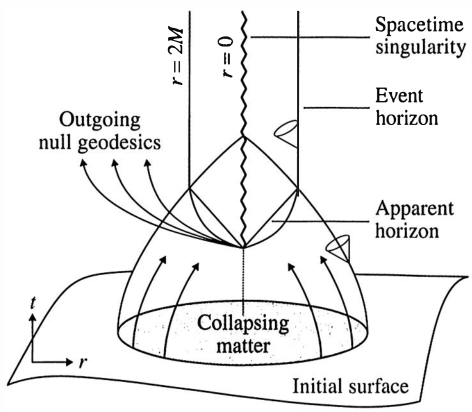


Fig. 20.2 The gravitational collapse of a spherical but inhomogeneous dust cloud with a density profile peaked at the center. The event horizon no longer forms prior to the singularity and the collapse end state is a naked singularity

of nakedness or otherwise of the singularity can be reduced to the existence of real, positive roots of an algebraic equation $V(X) = 0$, constructed out of the free functions F and f and their derivatives, which constitute the initial data of this problem. If the equation $V(X) = 0$ has a real positive root, the singularity could be naked. In order to be the end point of null geodesics at least one real positive value of X_0 should satisfy the above. If no real positive root of the above is found, the singularity is not naked. It should be noted that many real positive roots of the above equation may exist which give the possible values of tangents to the singular null geodesics terminating at the singularity in the past. Suppose now $X = X_0$ is a simple root to $V(X) = 0$. To determine whether X_0 is realized as a tangent along any outgoing singular geodesics to give a naked singularity, one can integrate the equation of the radial null geodesics in the form $r = r(X)$ and it is seen that there is always at least one null geodesic terminating at the singularity $t = 0, r = 0$, with $X = X_0$. In addition, there would be infinitely many integral curves as well, depending on the values of the parameters involved, which terminate at the singularity. It is thus seen that the existence of a positive real root of the equation $V(X) = 0$ is a necessary and sufficient condition for the singularity to be naked. Finally, to determine the curvature strength of the naked singularity at $t = 0$, $r = 0$, one may analyze the quantity $k^2 R_{ab} K^a K^b$ near the singularity. Standard analysis shows that the strong curvature condition is satisfied, in that the above quantity remains finite in the limit of approach to the singularity. The spacetime picture for a collapse terminating in a naked singularity is given in Fig. 20.2.

The assumption of vanishing pressures, which could be important in the final stages of the collapse, may be considered as the limitation of dust models. On the other hand, it is also argued sometimes that in the final stages of collapse the dust equation of state could be relevant and at higher and higher densities the matter may behave much more like dust. Further, if there are no large negative pressures (as implied by the validity of the energy conditions), then the pressure might also contribute gravitationally in a positive manner to the overall effect of dust and may not alter the final conclusions.

20.14 Collapse with General Matter Fields

It is clearly important to consider collapse situations which consider matter with nonzero pressures and with reasonable equations of state. It is possible that pressures may play an important role for the later stages of collapse and one must investigate the possibility whether pressure gradients can prevent the occurrence of naked singularity.

Many collapse scenarios have been considered by now with nonzero pressures and physically reasonable equations of state. What one needs to examine here again is the existence, the termination of future directed nonspace-like geodesic families at the singularity in the past, and the strength of such a singularity for collapse with nonzero pressure.

A useful insight into this issue is provided by self-similar collapse for a perfect fluid with a linear equation of state $p = k\rho$. A numerical treatment of self-similar perfect fluid collapse was given by *Ori* and *Piran* [20.32], and the analytic consideration for the same was given by *Joshi* and *Dwivedi* [20.33]. It can be seen that the collapse evolutions as allowed by the Einstein equations permit both the black hole and naked singularity final states. If in a self-similar collapse a single null radial geodesic escapes the singularity, then, in fact, an entire family of nonspace-like geodesics would also escape provided the positivity of energy density is satisfied. It also follows that no families of nonspace-like geodesics would escape the singularity, even though a single null trajectory might, if the weak energy condition is violated. The singularity will be globally visible to faraway observers in the spacetime for a wide set of conditions. These results show that naked singularity is not avoided by the introduction of nonzero pressures or a reasonable equation of state.

Actually, consideration of matter forms such as directed radiation, dust, perfect fluids, etc., imply that a similar general pattern emerges, as far as the final outcome of collapse is concerned. Basically, the result that emerges is that, depending on the nature of the regular initial data in terms of the density and pressure profiles, the Einstein equations permit both classes of dynamical evolutions, those leading to either of the black hole or naked singularity final states.

Hence one could ask the question whether the final fate of collapse would be independent of the form of the matter under consideration. An answer to this is useful

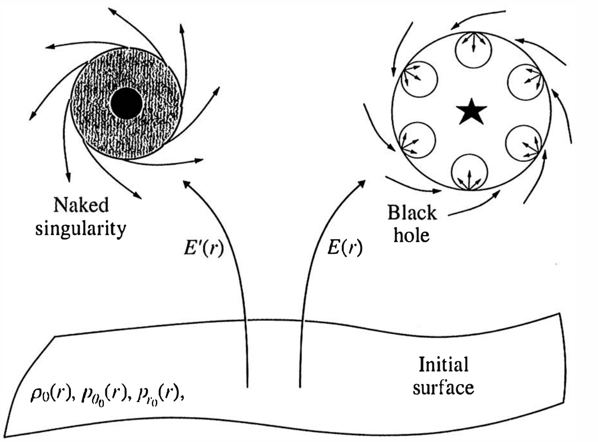


Fig. 20.3 For a generic collapse of a general matter field, the collapse final state can be either a black hole or a naked singularity depending on the dynamical evolution chosen as allowed by the Einstein equations

because it was often thought that once a suitable form of matter with an appropriate equation of state, also satisfying energy conditions, is considered then there may be no naked singularities. Of course, there is always a possibility that during the final stages of collapse the matter may not have any of the forms such as dust or perfect fluids considered above, because such relativistic fluids are phenomenological and perhaps one must treat matter in terms of some fundamental field, such as, for example, a massless scalar field. In that context, a naked singularity is also seen to form for the scalar field collapse [20.34], although for fine-tuned initial data.

In the above context, it is worth mentioning efforts in the direction of understanding collapse final states for general matter fields, which generalize the above results on perfect fluid to matter forms without any restriction on the form of T_{ij}, with the matter satisfying the weak energy condition. A consideration to a general form of matter was given by *Lake* [20.35] and by *Szekeres* and *Iyer* [20.36], who do not start by assuming an equation of state, but consider a class of metric coefficients with a certain power law behavior. Also, *Joshi* and *Dwivedi* [20.37] and *Goswami* and *Joshi* [20.38], and *Giambo* et al. [20.39] had results in this direction. The main argument is along the following lines. It was pointed out above that naked singularities could form in gravitational collapse from regular initial data from which nonzero measure families of nonspace-like trajectories come out. The criterion for the existence of such singularities was characterized in terms of the existence of real positive roots of an algebraic equation constructed out of the field variables. A similar procedure is developed now for the general form of matter. In comoving coordinates, the general matter can be described by three free functions, namely the energy density and radial and tangential pressures. The existence of naked singularity is again characterized in terms of the real positive roots of an algebraic equation, constructed from the equations of nonspace-like geodesics which involve the three metric functions. The field equations then relate these metric functions to the matter variables and it is seen that for a subspace of this free initial data in terms of matter variables, the above algebraic equation will have real positive roots, producing a naked singularity in the spacetime. When no such roots exist, the end state is a black hole.

It follows that the occurrence or otherwise of naked singularity is basically related to the choice of initial data to the Einstein field equations as determined by the evolutions allowed. Therefore, these occur from regular initial data within the general context considered, subject to the matter satisfying the weak energy condition. The condition on initial data which leads to the formation of black hole is also characterized.

It would then appear that the occurrence of naked singularity or a black hole is more a problem of the choice of the initial data for field equations rather than that of the form of matter or the equation of state (Fig. 20.3).

Such a conclusion has an important implication for cosmic censorship in that in order to preserve the same, one must avoid all such regular initial data causing naked singularity, and hence a much deeper understanding of the initial data space is required in order to determine such initial data and the kind of physical parameters they would specify. In other words, this classifies the range of physical parameters to be avoided for a particular form of matter. Such an understanding would also pave the way for black hole physics to use only those ranges of allowed parameter values which produce black holes only, thus putting black hole physics on a firmer footing.

20.15 Nonspherical Collapse and Numerical Simulations

Basically, the results and detailed studies such as the above on gravitational collapse show that cosmic censorship cannot hold in an unqualified general form. It must be properly fine-tuned and the black holes will form only under certain suitably restrictive conditions on collapse.

An important question at the same time is: what will be the final fate of gravitational collapse which is not spherically symmetric? The main phases of spherical collapse of a massive star would be typically instability, implosion of matter, and subsequent formation of an event horizon and a spacetime singularity of infinite density and curvature with infinite gravitational tidal forces. This singularity may or may not be fully covered by the horizon, as we discussed above.

As noted, small perturbations over the sphericity would leave the situation unchanged in the sense that an event horizon will continue to form in the advanced stages of the collapse. The next question then is, do horizons still form when the fluctuations from the spherical symmetry are high and the collapse is highly nonspherical? It was shown by *Thorne* [20.40], for example, that when there is no spherical symmetry, the collapse of infinite cylinders gives rise to naked singularities in general relativity, which are not covered by horizons. This situation motivated Thorne to propose the *hoop conjecture* for finite systems in an asymptotically flat spacetime for the final fate of nonspherical collapse. The horizons of gravity form when and only when a mass M gets compacted in a region whose circumference in *every* direction obeys $C \leq 2\pi(2GM/c^2)$. Thus, unlike cosmic censorship, the hoop conjecture does not rule out *all* naked singularities but only makes a definite assertion on the occurrence of event horizons in gravitational collapse. The hoop conjecture is concerned with the formation of event horizons and not with naked singularities. Thus, even when event horizons form, say, for example, in the spherically symmetric case, it does not rule out the existence of naked singularities, or it does not imply that such horizons must always cover the singularities.

When the collapse is sufficiently aspherical, with one or two dimensions being sufficiently larger than the others, the final state of collapse could be a naked singularity, according to the hoop conjecture. Such a situation is inspired by the *Lin* et al. [20.41] instability consideration in Newtonian gravity, where a nonrotating homogeneous spheroid collapses maintaining its homogeneity and spheroidicity but with growing deformations. If the initial condition is that of a slightly oblate spheroid, the collapse results in a pancake singularity through which the evolution could proceed. However, for a slightly prolate spheroidal configuration, the matter collapses to a thin thread which results into a spindle singularity. The gravitational potential and the tidal forces blow up as opposed to only density blowing up so it is a serious singularity. Even in the case of an oblate collapse, the passing of matter through the pancake causes prolateness and, subsequently, a spindle singularity again results without the formation of any horizon.

It is clear though that the nonspherical collapse scenario is rather complex to understand, and a recourse to the numerical simulations of evolving collapse models may greatly enhance our understanding of possible final collapse states in this case. In such a context, the numerical calculations of *Shapiro* and *Teukolsky* [20.42] indicated conformity with the hoop conjecture. They evolved collissionless gas spheroids in full general relativity, which in all cases collapse to singularities. When the spheroid is sufficiently compact a black hole may form, but when the semimajor axis of the spheroid is sufficiently large, a spindle singularity forms without the formation of an apparent horizon. This gives rise to the possibility of the occurrence of naked singularities in the collapse of finite systems in asymptotically flat spacetimes which violate weak cosmic censorship but are in accordance with the hoop conjecture.

We note that the Kerr black hole is believed to be the unique stationary solution in Einstein gravity when mass and rotation parameters are included. However, it is to be noted that while the Schwarzschild black hole is the final end state of homogeneous dust collapse, we have no interior solution for a rotating collapsing cloud. In other words, an exterior Kerr geometry has no internal solution in general relativity. We, therefore, do not really know the final fate of gravitational collapse with rotation. To understand the same, numerical simulations in full general relativity will be of great value. There are many such numerical programs in the making currently to deal with this problem of modeling a rotating collapsing massive star. The idea here is to include rotation in collapse and then to let the Einstein equations evolve the collapse to see if the Kerr black hole necessarily emerges as a final state (see, e.g., [20.43] and references therein). It is worth noting that numerical simulations in higher dimensions have also recently produced some very intriguing naked singularity formation scenarios (*Lehner* and *Pretorius* [20.44]).

We finally note that apart from such numerical simulations, certain analytic treatments of aspherical collapse are also available. For example, the nonspherical Szekeres models for irrotational dust without any Killing vectors, generalizing spherical Tolman–Bondi–Lemaître collapse, were studied by *Joshi* and *Królak* [20.45] to deduce the existence of strong curvature, naked singularities. While this indicates that naked singularities are not necessarily confined to spherical symmetry only, it is to be noted that dynamical evolution of a nonspherical collapse still remains a largely uncharted territory.

20.16 Are Naked Singularities Stable and Generic?

Naked singularities may develop in gravitational collapse, either spherical or otherwise. However, if they are not either generic or stable in some suitable sense, then they may not be necessarily physically relevant. An important question then is the genericity and stability of naked singularities arising from regular initial data. Will the initial data subspace, which gives rise to naked singularity as end state of collapse, have a vanishing measure in a suitable sense? In that case, one would be able to reformulate more suitably the censorship hypothesis, based on the criterion that naked singularities could form in collapse but may not be generic.

We note here that the *genericity* and *stability* of the collapse outcomes, in terms of black holes and naked singularities need to be understood carefully and in further detail. It is by and large well accepted now that the general theory of relativity allows and gives rise to both black holes and naked singularities as the final fate of a continual gravitational collapse, evolving from regular initial data and under reasonable physical conditions. What is not fully clear as yet is the distribution of these outcomes in the space of all allowed outcomes of collapse. The collapse models discussed above and the considerations we have given here would be of some help in this direction and may throw some light on the distribution of black holes and naked singularity solutions in the initial data space. For some considerations on this issue, especially in the context of scalar field collapse, we refer to *Christodoulou* [20.46] and *Joshi* et al. [20.47], and references therein, for further discussion. For the case of inhomogeneous dust collapse, the black hole and naked singularity spaces are shown in Fig. 20.4.

The important point, however, is that in general relativity there is no well-defined concept or formulation as to what to call generic and stable outcomes, unlike the Newtonian case. In other words, there are no well-defined criteria or definitions available as to what is meant by stability in general relativity. The ambiguity mainly arises because of nonunique topologies on the space of all Lorentzian metrics on a given spacetime manifold and a similar nonuniqueness of measures. In this situation there is no easy way to answer the question in any unique and definite manner, and people generally resort to the physical meaningfulness of the collapse scenario, which gives rise to either the black hole or the naked singularity outcome.

From such a perspective, it is natural and meaningful to ask here, what is really the physics that causes a naked singularity to develop in collapse, rather than a black hole? We need to know how at all particles and energy are allowed to escape from extremely strong gravity fields. We have examined this issue in some detail to bring out the role of inhomogeneities and spacetime shear towards distorting the geometry of horizons that form in collapse (Fig. 20.5).

In Newtonian gravity, it is only the matter density that determines the gravitational field. In Einstein the-

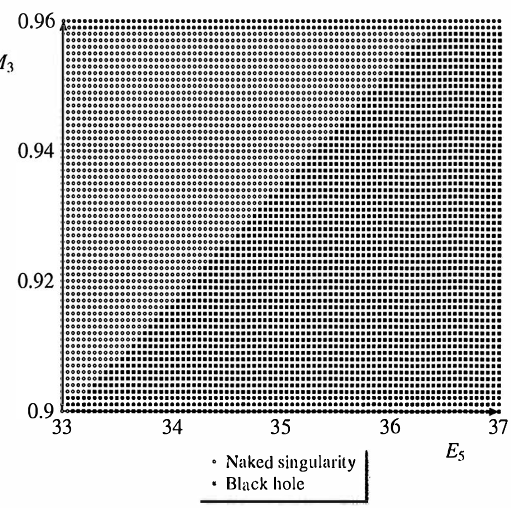


Fig. 20.4 Initial data leading to black holes and naked singularities shown in the spaces of mass and energy functions

Part C | 20.16

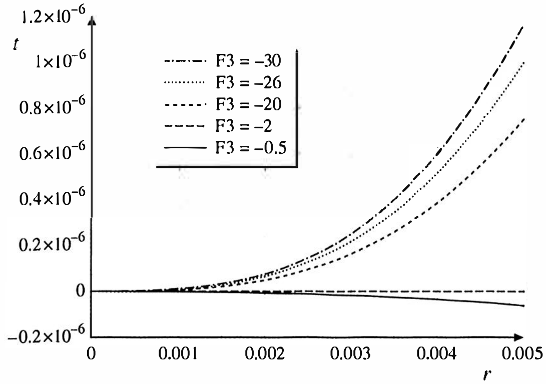


Fig. 20.5 The apparent horizon formation is delayed depending on the amount of inhomogeneity present as the collapse proceeds (after [20.48])

ory, however, density is only one attribute of the overall gravitational field, and the various curvature components and scalar quantities play an equally important role to dictate what the overall nature of the field is. What we have shown is that once the density is inhomogeneous or higher at the center of the collapsing star, this rather naturally delays the trapping of light and matter during collapse, which can in principle escape. This is a general relativistic effect to imply that even if the densities are very high, there are paths available for light or matter to escape due to inhomogeneously collapsing matter fields. These physical features then naturally lead to a naked singularity formation [20.48].

As it turns out, it is the amount of inhomogeneity that counts towards distorting the apparent horizon formation. If it is very small, below a critical limit, a black hole will form, but with sufficient inhomogeneity the trapping is delayed to cause a naked singularity. This criticality also comes out in the Vaidya class of radiation collapse models, where it is the rate of collapse, that is how fast or slow the cloud is collapsing, which determines the black hole or naked singularity formation.

20.17 Astrophysical and Observational Aspects

It is clear that the black hole and naked singularity outcomes of a complete gravitational collapse for a massive star are very different from each other physically and would have quite different observational signatures. In the naked singularity case, if it occurs in nature, we have the possibility to observe the physical effects occurring in the vicinity of the ultra-dense regions that form in the very final stages of collapse. However, in a black hole scenario, such regions are necessarily hidden within the event horizon of gravity.

There have been attempts where researchers explored physical applications and implications of the naked singularities (see, e.g., [20.49] and references therein). If we could find astrophysical applications of the models that predict naked singularities as collapse final fate and possibly try to test these through observational methods and the signatures predicted, this could offer a very interesting avenue to obtain further insight into the problem as a whole. An attractive recent possibility in this connection is to explore naked singularities as possible particle accelerators [20.50], where the possibility also emerges that the Cauchy horizons may not be innocuous, and high energy collisions could occur in the vicinity of the same if they are generated by naked singularity (Fig. 20.6).

Also, the accretion discs around a naked singularity, wherein the matter particles are attracted towards or repulsed away from the singularities with great velocities could provide an excellent venue to test such effects and may lead to predictions of important observational signatures to distinguish the black holes and naked singularities in astrophysical phenomena. It is then necessary to investigate the question of what observational

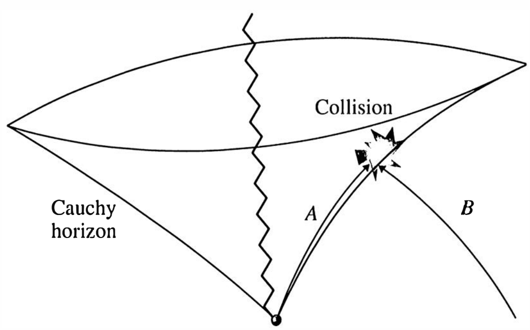


Fig. 20.6 Very high energy particle collisions can occur in the vicinity of the Cauchy horizon emerging from the naked singularity

Part C | 20.17

signatures would then emerge and to distinguish the black holes from naked singularities, and we must explore what special astrophysical consequences the latter may have.

One may ask several intriguing questions such as: *Where could the observational signatures of naked singularities lie?* If we look for the sign of singularities such as the ones that appear at the end of collapse, we must consider explosive and high energy events. In fact, such models expose the ultra-high density region at the time of the formation of the singularity while the outer shells are still falling towards the center. In such a case, shock waves emanating from the superdense region at scales smaller than the Schwarzschild radius (which could be due to quantum effects or repulsive classical effects) and collisions of particles near the Cauchy horizon could have effects on the outer layers. These would be considerably different from those appearing during the formation of a black hole. If, on the other hand, we consider singularities such as the super-spinning Kerr solution we can look for different kinds of observational signatures. Among these the most prominent features deal with the way the singularity may affect incoming particles, either in the form of light bending, such as in gravitational lensing, particle collisions close to the singularity, or properties of accretion disks.

Essentially what we ask is: *Could we test censorship using astronomical observations*? With so many high technology power missions to observe the cosmos, can we not just observe the skies carefully to determine the validity or otherwise of cosmic censorship? In this connection, several proposals to measure the mass and spin ratio for compact objects and for the galactic center have been made by different researchers. In particular, using pulsar observations it has been suggested that gravitational waves and the spectra of x-ray binaries could test the rotation parameter for the center of our galaxy. Also, the shadow cast by the compact object can be used to test the same in stellar mass objects, or the x-ray energy spectrum emitted by the accretion disk can be used. Using certain observable properties of gravitational lensing that depend upon rotation has also been suggested (for references, see [20.49]).

The basic issue here is that of sensitivity, namely how accurately and precisely we can measure and determine these parameters. A number of present and future astronomical missions may be of help. One of these is the Square-Kilometer Array (SKA) radio telescope, which will offer a possibility here, with a collecting area exceeding a factor of 100 compared to existing ones. SKA astronomers point out they will have the sensitivity desired to measure the required quantities to very precisely determine vital fundamental issues in gravitation physics, such as cosmic censorship, and to decide on its validity or otherwise. Other missions that could in principle provide a huge amount of observational data are those that are currently hunting for gravitational waves. Gravitational wave astronomy has yet to claim its first detection of waves, nevertheless in the coming years it is very likely that the first observations will be made by experiments such as LIGO and VIRGO, which are currently still below the threshold for observation. Then gravitational wave astronomy will become an active field with possibly large amounts of data to be checked against theoretical predictions, and it appears almost certain that this will have a strong impact on open theoretical issues such as the cosmic censorship problem.

There are three different kinds of observations that one could devise in order to distinguish a naked singularity from a black hole. The first one relies on the study of accretion disks. The accretion properties of particles falling onto a naked singularity would be very different from those of black hole of the same mass (see, for example, [20.51]), and the resulting accretion disks would also be observationally different. The properties of accretion disks have been studied in terms of the radiant energy, flux, and luminosity, in a Kerr-like geometry with a naked singularity, and the differences from a black hole accretion disk have been investigated. Also, the presence of a naked singularity gives rise to powerful repulsive forces that create an outflow of particles from the accretion disk on the equatorial plane. This outflow that is otherwise not present in the black hole case, could, in principle, be distinguished from the jets of particles that are thought to be ejected from a black hole's polar region and which are due to strong electromagnetic fields. Also, when charged test particles are considered the accretion disk's properties for the naked singularity present in the Reissner–Nordstrom spacetime are seen to be observationally different from those of black holes.

The second way of distinguishing black holes from naked singularities relies on gravitational lensing. It is argued that when the spacetime does not possess a photon sphere, the lensing features of light passing close to the singularity will be observationally different from those of a black hole. This method, however, does not appear to be very effective when a photon sphere is present in the spacetime. Assuming that a Kerr-like solution of Einstein equations with massless scalar field

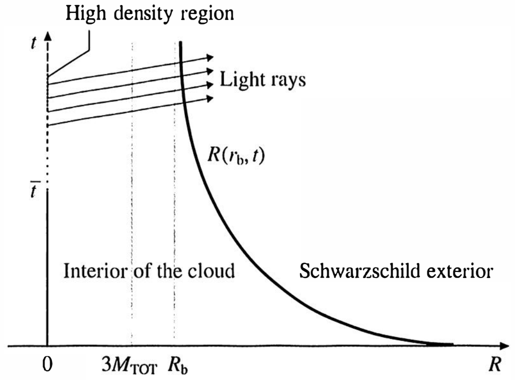


Fig. 20.7 An equilibrium configuration can be obtained from gravitational collapse which halts asymptotically and which would contain a central naked singularity

exists at the center of galaxies, its lensing properties are studied and it was found that there are effects due to the presence of both the rotation and scalar field that would affect the behavior of the bending angle of the light ray, thus making those objects observationally different from black holes.

Finally, a third way of distinguishing black holes from naked singularities comes from particle collisions and particle acceleration in the vicinity of the singularity. In fact, it is possible that the repulsive effects due to singularity can deviate a class of infalling particles, making these outgoing eventually. These could then collide with some ingoing particle, and the energy of collision could be arbitrarily high, depending on the impact parameter of the outgoing particle with respect to the singularity. The net effect is thus the creation of a very high energy collision that resembles that of an immense particle accelerator and that would be impossible in the vicinity of a Kerr black hole.

It was pointed our recently by *Joshi* et al. [20.52] that one could obtain equilibrium configurations as the final outcome of a gravitational collapse. If such an object arises without trapped surfaces but with a singularity at the center (see Fig. 20.7) then again the accretion disk properties are very different from a black hole of the same mass.

20.18 Predictability and Other Cosmic Puzzles

What then is the status of naked singularities versus censorship today? Can cosmic censorship survive in some limited and specialized form, and firstly, can we properly formulate it after all these studies on gravitational collapse in recent years? While this continues to be a major cosmic puzzle, recent studies on the formation of naked singularities as collapse end states for many realistic models have brought to forefront some of the most intriguing basic questions, both at classical and quantum levels, which may have significant physical relevance. Some of these are: can the super ultra-dense regions forming in a physically realistic collapse of a massive star be visible to far away observers in spacetime? Are there any observable astrophysical consequences? What is the causal structure of spacetime in the vicinity of singularity as decided by the internal dynamics of collapse which evolves from regular initial data at an initial time? How early or late will the horizons actually develop in a physically realistic gravitational collapse, as determined by the astrophysical conditions within the star? When a naked singularity forms, is it possible to observe the quantum gravity effects taking place in the ultra-strong gravity regions? Can one possibly envisage a connection to observed ultra-high energy phenomena such as cosmic gamma-ray bursts?

A continuing study of collapse phenomena within a general and physically realistic framework may be the only way to find answers to some of these issues. This could lead us to novel physical insights and possibilities emerging out of the intricacies of gravitational force and the nature of gravity, as emerging from examining the dynamical evolutions as allowed by Einstein equations.

Apart from their physical relevance, the collapse phenomena also have profound philosophical implications such as on the issue of predictability in the universe. Below, we summarize a few arguments for and against this in classical general relativity.

It is sometimes argued that the breakdown of censorship means violation of predictability in spacetime, because we have no direct handle to know what a naked singularity may radiate and emit unless we study the physics in such ultra-dense regions. One would then not be able to predict the universe in the future of a given epoch of time as would be the case, for example, in the case of the Schwarzschild black hole that develops in Oppenheimer–Snyder collapse. A concern that

is usually expressed is if naked singularities occurred as the final fate of gravitational collapse, predictability is violated in spacetime because naked singularity is characterized by the existence of light rays and particles that emerge from the same. Typically, in all the collapse models discussed above, there is a family of future directed nonspace-like curves that reach external observers, and when extended in the past they met the singularity. The first light ray that comes out from the singularity marks the boundary of the region that can be predicted from a regular initial Cauchy surface in the spacetime, and this is called the Cauchy horizon for the spacetime. The causal structure of spacetime would differ significantly in the two cases when there is a Cauchy horizon and when there is none.

In general relativity, a given *epoch* of time is sometimes represented by a space-like surface, which is a three-dimensional space section. For example, in the standard Friedmann models of cosmology, there is such an epoch of simultaneity, from which the universe evolves in future, given the physical variables and initial data on this surface. The Einstein equations govern this evolution of universe, and there is thus a predictability which one would expect to hold in a classical theory. The concern then is that one would not be able to predict the future of naked singularity, and that unpredictable inputs may emerge from the same.

Given regular initial data on a space-like hypersurface, one would like to predict the future and past evolutions in spacetime for all times (see, for example, [20.1]). Such a requirement is termed the *global hyperbolicity* of the spacetime. A globally hyperbolic spacetime is a fully predictable universe; it admits a *Cauchy surface*, which is a three-dimensional space-like surface, the data on which can be evolved for all times in the past as well as in the future. Simple enough spacetimes such as Minkowski or Schwarzschild are globally hyperbolic, but Reissner–Nordstrom or Kerr geometries are not globally hyperbolic. For further details on these issues, we refer to [20.26].

The key role that the event horizon of a black hole plays is that it hides the super ultra-dense region formed in collapse from us. So the fact that we do not understand such regions has no effect on our ability to predict what happens in the universe at large. However, if no such horizon exists, then the ultra-dense region might, in fact, play an important and even decisive role in the rest of the universe, and our ignorance of such regions would become of more than merely academic interest.

Yet such an unpredictability is common in general relativity and not always directly related to censorship violation. Even black holes themselves need not fully respect predictability when they rotate or have some charge. For example, if we drop an electric charge into an uncharged black hole, the spacetime geometry changes radically and is no longer predictable from a regular initial epoch of time. A charged black hole admits a naked singularity that is visible to an observer within the horizon, and a similar situation holds when the black hole is rotating. There has been an important debate in recent years whether one could over-charge or over-rotate a black hole so that the singularity visible to observers within the horizon would also become visible to external far away observers (see, e.g., [20.53]).

Also, if such a black hole were big enough on a cosmological scale, the observer within the horizon could, in principle, survive happily for millions of years without actually falling into the singularity, and would thus be able to observe the naked singularity for a long time. Thus, only the purest of pure black holes with no charge or rotation at all respects the full predictability, and all other physically realistic ones with charge or rotation actually do not. As such, there are many models of the universe in cosmology and relativity that are not totally predictable from a given space-like hypersurface in the past. In these universes, the spacetime cannot be neatly separated into space and time foliation so as to allow initial data at a given moment of time to fully determine the future.

Actually the real breakdown of predictability is the occurrence of spacetime singularity itself we could say, which indicates the true limitation of the classical gravity theory. It does not matter really whether or not it is hidden within an event horizon. The real solution of the problem would then be the resolution of singularity itself, through either a quantum theory of gravity or in some way at the classical level.

In fact, the cosmic censorship way to predictability, that of *hiding the singularity within a black hole*, and then thinking that we restored the spacetime predictability may not be the real solution, or at best it may be only a partial solution to the key issue of predictability in spacetime universes. It may be just shifting the problem elsewhere, and some of the current major paradoxes faced by black hole physics such as the information paradox, the various puzzles regarding the nature of Hawking radiation, and other issues could also be a manifestation of the same.

No doubt, the biggest argument in support of censorship is that it would justify and validate the extensive formalism and laws of black hole physics and the astrophysical applications made so far. Censorship has

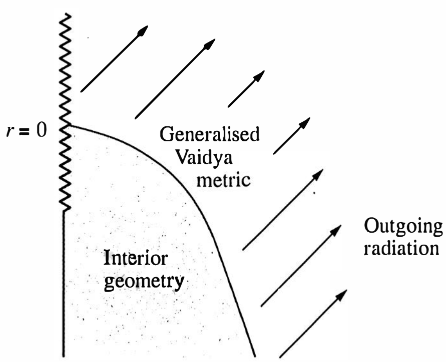


Fig. 20.8 If the star could radiate away very considerable mass, especially through negative quantum pressures close to the classical singularity, this may effectively resolve the singularity

been the foundation for the laws of black holes such as the area theorem and others, and their astrophysical applications. However, these are not free of major paradoxes. Even if we accept that all massive stars would necessarily turn into black holes, this still creates some major physical paradoxes. Firstly, all the matter entering a black hole must of necessity collapse into a spacetime singularity of infinite density and curvatures, where all known laws of physics break down, which is some kind of instability at the classical level itself. This was a reason why many gravitation theorists of the 1940s and the 1950s objected to black hole formation, and Einstein also repeatedly argued against such a final fate of a collapsing star, writing a paper in 1939 to this effect. Also, as is well known and has been widely discussed in the past few years, a black hole, by potentially destroying information, appears to contradict the basic principles of quantum theory. In that sense, the very formation of a black hole itself with a singularity within it appears to come laden with inherent problems. It is far from clear how one would resolve these basic troubles even if censorship were correct.

In view of such problems with the black hole paradigm, a possibility worth considering is the delay or avoidance of horizon formation as the star collapses under gravity. This happens when collapse to a naked singularity takes place, namely, where the horizon does not form early enough or is avoided. In such a case, if the star could radiate away most of its mass in the late stages of collapse, this may offer a way out of the black hole conundrum, while also resolving the singularity issue, because now there is no mass left to form the singularity. While this may be difficult to achieve purely classically, such a phenomenon could happen when quantum gravity effects are taken into account (Fig. 20.8, see also the next section for a further discussion).

What this means is that such an *unpredictability* is somewhat common in general relativity. For example, if we drop a slight charge in a Schwarzschild black hole, the spacetime geometry completely changes into that of a charged black hole that is no longer predictable in the above sense. A similar situation holds when the black hole is rotating. In fact, there are very many models of universe in use in relativity that are not *globally hyperbolic*, that is, not totally predictable in the above sense where space and time are neatly separated so as to allow initial data to fully determine the future for all times.

In any case, a positive and useful feature that has emerged from work on collapse models so far is, we already now have several important constraints for any possible formulation of censorship. It can be seen that several versions of censorship proposed earlier would not hold, because explicit counter-examples are now available. Clearly, analyzing gravitational collapse plays a crucial role here. Only if we understand clearly why naked singularities develop as collapse end states in many realistic models, could there emerge any pointer or lead to any practical and provable version of censorship.

Finally, it may be worth noting that even if the problem of singularity were resolved somehow, possibly by invoking quantum gravity which may smear the singularity, we still have to mathematically formulate and prove the black hole formation assuming an appropriate censorship principle, which is turning out to be a most difficult task with no sign of resolution. As has been discussed, the detailed collapse calculations of recent years show that the final fate of a collapsing star could be a naked singularity in violation of censorship. Finally, as is well known and has been widely discussed by now, a black hole creates an information loss paradox, violating unitarity and contradicting basic principles of quantum theory. It is far from clear how one would resolve these basic troubles even if censorship were correct.

20.19 A Lab for Quantum Gravity–Quantum Stars?

It is believed that when we have a reasonable and complete quantum theory of gravity, all spacetime singularities, whether naked or those hidden inside black holes, will be resolved. As of now, it remains an open question whether quantum gravity will remove naked singularities. After all, the occurrence of spacetime singularities could be a purely classical phenomenon and whether they are naked or covered should not be relevant, because quantum gravity will possibly remove them all anyway. It is possible that in a suitable quantum gravity theory the singularities will be smeared out, although this has not been realized so far.

In any case, the important and real issue is whether or not the extreme strong gravity regions formed due to gravitational collapse are visible to faraway observers. It is quite clear that gravitational collapse would certainly proceed classically, at least until quantum gravity starts governing and dominating the dynamical evolution at scales of the order of the Planck length, i. e., until the extreme gravity configurations have already been developed due to collapse. The point is that it is the visibility or otherwise of such ultra-dense regions that is under discussion, whether they are classical or quantum (see Fig. 20.9).

What is important is that classical gravity necessarily implies the existence of ultra-strong gravity regions, where both classical and quantum gravity come into their own. In fact, if naked singularities develop in gravitational collapse, then in a literal sense we come face-to-face with the laws of quantum gravity whenever such an event occurs in the universe.

In this way, the gravitational collapse phenomenon has the potential to provide us with a possibility of actually testing the laws of quantum gravity. In the case of a black hole developing in the collapse of a finite sized object such as a massive star, such strong gravity regions are necessarily hidden behind an event horizon of gravity, and this would be well before the physical conditions become extreme near the spacetime singularity. In that case, the quantum effects, even if they cause qualitative changes closer to singularity, will be of no physical consequence as no causal communications are then allowed from such regions. On the other hand, if the causal structure were that of a naked singularity, then the communications from such a quantum gravity dominated extreme curvature ball would be visible in principle. This would be so either through direct physical processes near a strong curvature naked singularity, or via the secondary effects, such as the shocks produced in the surrounding medium. It is possible that a spacetime singularity basically represents the incompleteness of the classical theory and when quantum effects are combined with the gravitational force, the classical singularity may be resolved.

Therefore, more than the existence of a naked singularity, the important physical issue is whether the extreme gravity regions formed in the gravitational collapse of a massive star are visible to external observers in the universe. An affirmative answer here would mean that such a collapse provides a good laboratory to study quantum gravity effects in the cosmos, which may possibly generate clues for an as yet unknown theory of quantum gravity. Quantum gravity theories in the making, such as string theory or loop quantum gravity, in fact, are badly in need of some kind of an observational input, without which it is nearly impossible to constrain the plethora of possibilities.

We could say quite realistically that a laboratory similar to that provided by the early universe is created in the collapse of a massive star. However, the big bang, which is also a naked singularity in that it is, in principle, visible to all observers, occurred only once in the life of the universe and is, therefore, a unique event. However, a naked singularity of gravitational collapse could offer an opportunity to explore and observe the quantum gravity effects every time a massive star in the universe ends its life.

The important questions one could ask are the following: If in realistic astrophysical situations the star terminates as a naked singularity, would there be any observable consequences which reflect the quantum gravity signatures in the ultra-strong gravity region? Do naked singularities have physical properties different from those of a black hole? Such questions underlie our study of gravitational collapse.

In view of recent results on gravitational collapse, and various problems with the black hole paradigm, a possibility worth considering is the delay or avoidance of horizon formation as the star evolves collapsing under gravity. This happens when collapse to a naked singularity takes place, where the horizon does not form early enough or is avoided. In such a case, in the late stages of collapse if the star could radiate away most of its mass, then this may offer a way out of the black hole conundrum, while also resolving the singularity issue, because now there is no mass left to form the curvature singularity. The purpose is to resolve the black hole paradoxes and avoid the singularity, either vis-

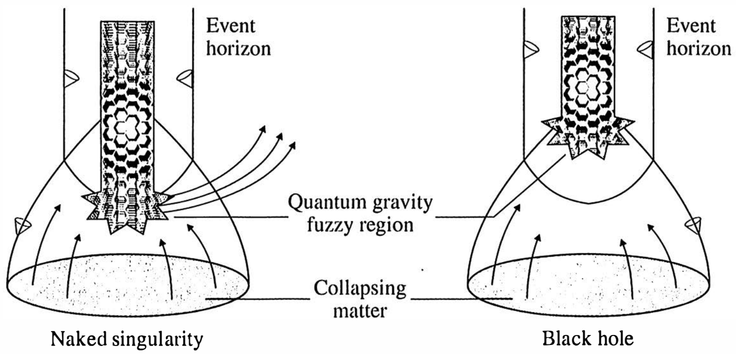


Fig. 20.9 Naked singularity may be resolved by quantum gravity effects but the ultra-strong gravity region that developed in gravitational collapse will still be visible to external observers in the universe

ible or within a black hole, which actually indicates the breakdown of physical theory. The current work on gravitational collapse suggests possibilities in this direction.

In this context, we considered a cloud that collapsed to a naked singularity final state, and introduced loop quantum gravity effects [20.54] (see also [20.55]). It turned out that the quantum effects generated an extremely powerful repulsive force within the cloud. Classically the cloud would have terminated into a naked singularity, but quantum effects caused a burst-like emission of matter in the very last phases of collapse, thus dispersing the star and dissolving the naked singularity. The density remained finite and the spacetime singularity was eventually avoided. One could expect this to be a fundamental feature of other quantum gravity theories as well, but more work would be required to confirm such a conjecture.

For a realistic star, its final catastrophic collapse takes place in a matter of seconds. A star that lived millions of years thus collapses in only tens of seconds. In the very last fraction of a microsecond, almost a quarter of its total mass must be emitted due to quantum effects, and, therefore, this would appear like a massive, abrupt burst to an external observer far away. Typically, such a burst will also carry with it specific signatures of quantum effects taking place in such ultra-dense regions. In our case, these included a sudden dip in the intensity of emission just before the final burst-like evaporation due to quantum gravity. The question is, whether such unique astrophysical signatures can be detected by modern experiments, and if so, what they tell us about quantum gravity, and if there are any new insights into other aspects of cosmology and fundamental theories such as string theory. The key point is that because the very final ultra-dense regions of the star are no longer hidden within a horizon as in the black hole case, the exciting possibility of observing these quantum effects now arises, independently of the quantum gravity theory used. An astrophysical connection to extreme high energy phenomena in the universe, such as the gamma-ray bursts that have defied any explanations so far, may not be ruled out.

Such a resolution of naked singularity through quantum gravity could be a solution to some of the paradoxes mentioned above (see also [20.56]; for other possibilities on singularity resolution). Then, whenever a massive star undergoes a gravitational collapse, this might create a laboratory for quantum gravity in the form of a *quantum star* (see, e.g., [20.53]), that we may possibly be able to access. This would also suggest intriguing connections to high energy astrophysical phenomena. The present situation poses one of the most interesting challenges that has emerged through the recent work on gravitational collapse.

We hope the considerations here have shown that gravitational collapse, which is essentially the investigation of dynamical evolutions of matter fields under the force of gravity in spacetime, provides one of the most exciting research frontiers in gravitation physics and high energy astrophysics. In our view, there is scope, therefore, for both theoretical as well as numerical investigations in these areas, which may have much to tell for our quest on basic issues in quantum gravity, fundamental physics, and gravity theories, and towards the expanding frontiers of modern high energy astrophysical observations.

20.20 Concluding Remarks

We have considered here several aspects of spacetime singularities and the physical scenarios where these may be relevant because they play an interesting and intriguing role. We hope this creates a fairly good view of the exciting new physics that the spacetime singularities are leading us to, presenting a whole spectrum of new possibilities towards our search of the universe.

After discussing their existence and certain key basic properties, we discussed in some detail the gravitational collapse scenarios and the useful conclusions that have emerged so far in this context. In the first place, singularities not covered fully by the event horizon occur in several collapsing configurations from regular initial data, with reasonable equations of state such as those describing radiation, dust, or a perfect fluid with a nonzero pressure, or also for general forms of matter. These naked singularities are physically significant in that densities and curvatures diverge powerfully near the same. Such results on the final fate of collapse, generated from the study of different physically reasonable collapse scenarios, may provide useful insights into black hole physics and may be of help for any possible formulation of the cosmic censorship hypothesis.

An insight that seems to emerge is that the final state of a collapsing star, in terms of either a black hole or a naked singularity, may not really depend on the form or equation of state of collapsing matter, but is actually determined by the physical initial data in terms of the initial density profiles and pressures.

As an example, for inhomogeneous dust collapse, the final fate could be a black hole or a naked singularity depending on the values of initial parameters. The collapse ends in a naked singularity if the leading nonvanishing derivative of density at the center is either the first one or the second one. There is a transition from the naked singularity phase to the black hole phase as the initial density profile is made more and more homogeneous near the center. As one progresses towards more homogeneity, and hence towards a stronger gravitational field, there first occurs a weak naked singularity, then a strong naked singularity, and finally a black hole.

The important question then is the genericity and stability of such naked singularities arising from regular initial data. Will the initial data subspace giving rise to naked singularity have zero measure in a suitable sense? In that case, one would be able to reformulate more suitably the censorship hypothesis, based on a criterion that naked singularities could form in collapse but may not be generic. As we pointed out, the answer is far from clear due to ambiguities in the definitions of measures and the stability criteria.

One may try to evolve some kind of a physical formulation for cosmic censorship, where the available studies on various gravitational collapse scenarios such as the above may provide a useful guide. The various properties of naked singularities may be collectively studied as they emerge from the studies so far, and one would then argue that objects with such properties are not physical. However, the way forward is again far from clear.

One could also invoke quantum effects and quantum gravity. While naked singularities may form in classical general relativity, quantum gravity presumably removes them. The point is that even though the final singularity may be removed in this way, there would still be very high density and curvature regions in the classical regime which would be causally communicating with outside observers, as opposed to the black hole case. If quantum effects could remove the naked singularity, this would then be some kind of *quantum cosmic censorship*.

We hope the considerations here have shown that gravitational collapse, which essentially is the investigation of dynamical evolutions of matter fields under the force of gravity in spacetime, provides one of the most exciting research frontiers in gravitation physics and high energy astrophysics. There are issues here which have deep relevance both for theory as well as observational aspects in astrophysics and cosmology. Also these problems are of relevance for the basics of gravitation theory and quantum gravity, and these inspire a philosophical interest and inquiry into the nature and structure of spacetime, causality, and profound issues such as predictability in the universe, as we have indicated here.

Research is already taking place in many of these areas as the discussion here has pointed out. Some of the most interesting questions from the author's personal perspective are: genericity and stability of collapse outcomes, examination of the quantum gravity effects near singularities, observational and astrophysical signatures of the collapse outcomes, and other related issues. In particular, one of the most interesting questions would be, if naked singularities which are hypothetical astrophysical objects, did actually form in nature, what distinct observational signatures would they present? That is, how one distinguishes black holes

from naked singularities would be an important issue. There have been some efforts on this issue in recent years, as we indicated above. The point is that there are already very high energy astrophysical phenomena being observed today, with several observational missions working both from ground and space. The black holes and naked singularities, which are logical consequences of star collapse in general relativity, would appear to be the leading candidates to explain these phenomena. The observational signatures that each of these would present, and their astrophysical consequences would be of much interest for future theoretical and computational research, and for their astrophysical applications.

In our view, there is, therefore, a scope for both theoretical as well as numerical investigations in these frontier areas, which may have much to say for our quest on basic issues in quantum gravity, fundamental physics, and gravity theories, and towards the expanding frontiers of modern high energy astrophysical observations.

References

20.1 S.W. Hawking, G.F.R. Ellis: *The Large Scale Structure of Spacetime* (Cambridge Univ. Press, Cambridge 1973)

20.2 R. Wald: *General Relativity* (Univ. Chicago Press, Chicago 1984)

20.3 G.F.R. Ellis, A. King: Was the big bang a whimper?, Commun. Math. Phys. **38**, 119 (1974)

20.4 G.F.R. Ellis, B. Schmidt: Singular spacetimes, Gen. Relativ. Gravit. **8**, 915 (1977)

20.5 A.K. Raychaudhuri: Relativistic cosmology, Phys. Rev. **98**, 1123 (1955)

20.6 V.A. Belinskii, I.M. Khalatnikov, E.M. Lifshitz: A general solution of the Einstein equations with a time singularity, Adv. Phys. **31**, 639–667 (1982)

20.7 C. Uggla, H. van Elst, J. Wainwright, G.F.R. Ellis: Past attractor in inhomogeneous cosmology, Phys. Rev. D **68**, 1–22 (2003)

20.8 C. Uggla: Recent developments concerning generic spacelike singularities, Gen. Relativ. Gravit. **45**, 1669–1710 (2013)

20.9 S.W. Hawking, R. Penrose: The singularities of gravitational collapse and cosmology, Proc. R. Soc. A **314**, 529 (1970)

20.10 P.S. Joshi: *Global Aspects in Gravitation and Cosmology* (Oxford Univ. Press, Oxford 1993)

20.11 C.J.S. Clarke: Singularities: Global and local aspects. In: *Topological Properties and Global Structure of Space-time*, ed. by P.G. Bergmann, V. de Sabbata (Plenum, New York 1986)

20.12 C.J.S. Clarke: Singularities in globally hyperbolic spacetimes, Commun. Math. Phys. **41**, 65 (1975)

20.13 F. Tipler: Singularities in conformally flat spacetimes, Phys. Lett. A **64**, 8 (1977)

20.14 F. Tipler, C.J.S. Clarke, G.F.R. Ellis: Singularities and horizons. In: *General Relativity and Gravitation*, Vol. 2, ed. by A. Held (Plenum, New York 1980) pp. 97–206

20.15 C.J.S. Clarke, A. Królak: Conditions for the occurrence of strong curvature singularities, J. Geo. Phys. **2**, 127 (1986)

20.16 R. Schoen, S.-T. Yau: The existence of a black hole due to condensation of matter, Commun. Math. Phys. **90**, 575 (1983)

20.17 K. Gödel: An example of a new type of cosmological solution of Einsteins field equations of gravitation, Rev. Mod. Phys. **21**, 447 (1949)

20.18 F. Tipler: Causality violations in asymptotically flat spacetimes, Phys. Rev. Lett. **37**, 879 (1976)

20.19 E. Minguzzi: Chronological spacetimes without lightlike lines are stably causal, Commun. Math. Phys. **288**, 801–819 (2009)

20.20 P.S. Joshi: On higher order causality violations, Phys. Lett. A **85**, 319 (1981)

20.21 P.S. Joshi, R.V. Saraykar: Cosmic censorship and topology change in general relativity, Phys. Lett. A **120**, 111 (1987)

20.22 C.J.S. Clarke, P.S. Joshi: On reflecting spacetimes, Class. Quantum Gravity **5**, 19 (1988)

20.23 M. Kriele: Causality violations and singularities, Gen. Relativ. Gravit. **22**, 619 (1990)

20.24 C.J.S. Clarke, F. de Felice: Globally non-causal spacetimes, J. Phys. A **15**, 2415 (1982)

20.25 J.M.M. Senovilla: New class of inhomogeneous cosmological perfect-fluid solutions without big-bang singularity, Phys. Rev. Lett. **64**, 2219 (1990)

20.26 P.S. Joshi: *Gravitational Collapse and Spacetime Singularities* (Cambridge Univ. Press, Cambridge 2008)

20.27 J.R. Oppenheimer, H. Snyder: On continued gravitational contraction, Phys. Rev. **56**, 455 (1939)

20.28 B. Datt: Über eine Klasse von Lösungen der Gravitationsgleichungen der Relativität, Z. Phys. **108**, 314 (1938)

20.29 R. Penrose: Gravitational collapse: the role of general relativity, Riv. Nuovo Cimento (Numero Speziale) I, 257–276 (1969)

20.30 R. Penrose: Singularities and time asymmetry. In: *General Relativity – an Einstein Centenary Survey*, ed. by S.W. Hawking, W. Israel (Cambridge Univ. Press, Cambridge 1979)

20.31 P.S. Joshi, I.H. Dwivedi: Naked singularities in spherically symmetric inhomogeneous Tolman–Bondi dust cloud collapse, Phys. Rev. D **47**, 5357 (1993)
20.32 A. Ori, T. Piran: Naked singularity in self-similar spherical gravitational collapse, Phys. Rev. Lett. **59**, 2137 (1987)
20.33 P.S. Joshi, I.H. Dwivedi: The Structure of naked singularity in self-similar gravitational collapse, Commun. Math. Phys. **146**, 333 (1992)
20.34 M.W. Choptuik: Universality and scaling in gravitational collapse of a massless scalar field, Phys. Rev. Lett. **70**, 9 (1993)
20.35 K. Lake: Precursory singularities in spherical gravitational collapse, Phys. Rev. Lett. **68**, 3129 (1992)
20.36 P. Szekeres, V. Iyer: Spherically symmetric singularities and strong cosmic censorship, Phys. Rev. D **47**, 4362 (1993)
20.37 P.S. Joshi, I.H. Dwivedi: Initial data and the end state of spherically symmetric gravitational collapse, Class. Quantum Gravity **16**, 41 (1999)
20.38 R. Goswami, P.S. Joshi: Spherical gravitational collapse in *N*-dimensions, Phys. Rev. D **76**, 084026 (2007)
20.39 R. Giambo, F. Giannoni, G. Magli, P. Piccione: Naked singularities formation in the gravitational collapse of barotropic spherical fluids, Gen. Relativ. Gravit. **36**, 1279 (2004)
20.40 K.S. Thorne: Non-spherical gravitational collapse: A short review. In: *Magic without Magic – John Archibald Wheeler*, ed. by J. Clauder (Freeman, New York 1972)
20.41 C.C. Lin, L. Mestel, F.H. Shu: The gravitational collapse of a uniform spheroid, Astrophys. J. **142**, 1431 (1965)
20.42 S.L. Shapiro, S.A. Teukolsky: Formation of naked singularities: The violation of cosmic censorship, Phys. Rev. Lett. **66**, 994 (1991)
20.43 B. Giacomazzo, L. Rezzolla, N. Stergioulas: Collapse of differentially rotating neutron stars and cosmic censorship, Phys. Rev. D **84**, 024022 (2011)
20.44 L. Lehner, F. Pretorius: Black strings, low viscosity fluids, and violation of cosmic censorship, Phys. Rev. Lett. **105**, 101102 (2010)
20.45 P.S. Joshi, A. Królak: Naked strong curvature singularities in Szekeres spacetimes, Class. Quantum Gravity **13**, 3069 (1996)
20.46 D. Christodoulou: The instability of naked singularities in the gravitational collapse of a scalar field, Ann. Math. **149**, 183 (1999)
20.47 P.S. Joshi, D. Malafarina, R.V. Saraykar: Genericity aspects in gravitational collapse to black holes and naked singularities, Int. J. Mod. Phys. D **21**, 1250066 (2012)
20.48 P.S. Joshi, N. Dadhich, R. Maartens: Why do naked singularities form in gravitational collapse?, Phys. Rev. D **65**, 101501 (2002)
20.49 P.S. Joshi, D. Malafarina: Recent developments in gravitational collapse and spacetime singularities, Int. J. Mod. Phys. D **20**, 2641 (2011)
20.50 M. Patil, P.S. Joshi: Kerr naked singularities as particle accelerators, Class. Quantum Gravity **28**, 235012 (2011)
20.51 D. Pugliese, H. Quevedo, R. Ruffini: Motion of charged test particles in Reissner–Nordstrom spacetime, Phys. Rev. D **83**, 104052 (2011)
20.52 P.S. Joshi, D. Malafarina, R. Narayan: Equilibrium configurations from gravitational collapse, Class. Quantum Gravity **28**, 235018 (2011)
20.53 P.S. Joshi: Naked singularities, Sci. Am. **300**, 36 (2009)
20.54 R. Goswami, P.S. Joshi, P. Singh: Quantum evaporation of a naked singularity, Phys. Rev. Lett. **96**, 031302 (2006)
20.55 T. Harada, H. Iguchi, K. Nakao: Naked singularity explosion, Phys. Rev. D **61**, 101502 (2000)
20.56 T. Biswas, E. Gerwick, T. Koivisto, A. Mazumdar: Towards singulary and ghost free theories of gravity, Phys. Rev. Lett. **108**, 031101 (2012)

21. Singularities in Cosmological Spacetimes

Beverly K. Berger

Theorems state that gravitational collapse from generic but non-singular initial conditions results in some type of singular behavior. Here the nature of the resultant approach to the singularity is examined in spatially homogeneous, anisotropic, vacuum cosmological spacetimes. The approach to the singularity in these spacetimes is either (asymptotically) Kasner-like or Mixmaster-like. It has been conjectured that spatially inhomogeneous cosmological spacetimes approach the singularity through Kasner-like or Mixmaster-like dynamics at every spatial point. Several examples of such cosmologies are explored numerically and heuristically. The current status of a rigorous statement of this conjecture and possible approaches to a proof are discussed. This chapter will focus on singularities in cosmological spacetimes.

21.1 Basic Concepts

21.1.1 Overview

Rather than reserve the word *cosmology* only for models to describe the actual universe, we shall generalize the concept to include spacetimes that are solutions to Einstein's equations and share the properties that distinguish a universe from, say, a binary black hole system. In the former, the matter is everywhere rather than localized. In addition, there is no asymptotically flat region outside the matter where, e.g., the binary's mass may be measured. Most of the following discussion will explore singularities in cosmological spacetimes that bear no relation to the actual universe. However, these spacetimes will provide theoretical laboratories to study issues related to singularities arising in generic, physical spacetimes.

To illustrate the meaning of singularities as they arise in cosmological spacetimes, we first consider a simple example. Spherical Bessel functions of order zero are defined to be solutions to

$$\frac{\mathrm{d}^2y}{\mathrm{d}t^2}+\frac{2}{t}\frac{\mathrm{d}y}{\mathrm{d}t}+y=0\,, \tag{21.1}$$

with the general solution

$$y=A\frac{\cos t}{t}+B\frac{\sin t}{t}\,, \tag{21.2}$$

for A and B arbitrary constants. Taylor expansions of $\cos t$ and $\sin t$ are used to show that the first term on the right-hand side of (21.2) $\to \pm\infty$ as $t\to 0$ while the second term has the finite value B. Figure 21.1 shows the structure of this singularity in the solutions to (21.1).

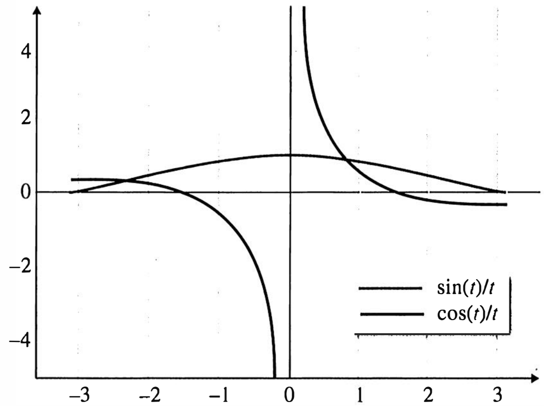


Fig. 21.1 Example of a singularity. The functions $\sin(t)/t$ (*brown*) and $\cos(t)/t$ (*black*) are shown for t in the interval $[-\pi, \pi]$

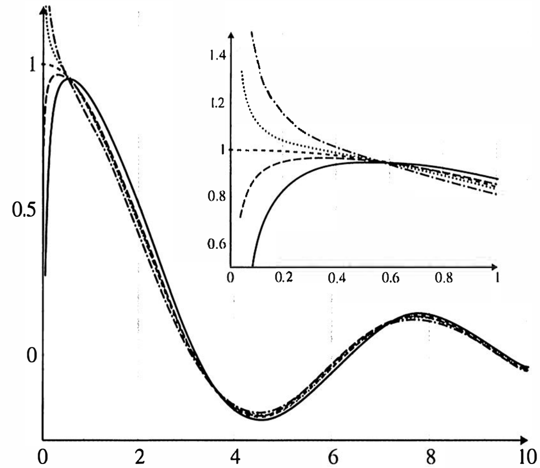


Fig. 21.2 Numerical search for a nongeneric solution. The *horizontal axis* is t and the *vertical axis* $y(t)$. Equation (21.1) is integrated from $y = y_0$ at $t = 10$ in the direction of $t = 0$, i.e., backwards in time. The initial velocity $\mathrm{d}y/\mathrm{d}t$ is adjusted to approach the nongeneric solution proportional to $\sin(t)/t$. The *inset* is an enlargement of the region near $t = 0$

In a sense, the singularity at $t = 0$ represents a boundary to the evolution of solutions to (21.1), since a solution starting at $t_0 > 0$ ($t_0 < 0$) cannot reach $t < 0$ ($t > 0$). We further notice that any solution to (21.1) with $A \neq 0$ has a singularity at $t = 0$. Such a solution is called *generic*. The special, nonsingular, solutions with $A = 0$ are said to be *nongeneric* since they represent a one-dimensional set within the two-dimensional parameter space defined by A and B. This simple example also illustrates that it is possible to use numerical methods to study singularities. Numerical evolution of, say, (21.1) cannot proceed if any simulation variable becomes infinite. Yet, it is possible to find the nongeneric solution numerically as is shown in Fig. 21.2. Here one takes advantage of the sign change in $\cos(t)/t$ as it approaches the singularity to zoom in on the special case $\sin(t)/t$.

Singularities are of more than mathematical interest when they occur in the evolution of equations for physical systems that start from well-behaved initial conditions. In principle, a singularity represents a breakdown in the equations that produce it. In the two examples from physics that follow, singularities do not present fundamental issues because they can be removed by replacing the singular equations with different ones that describe the physics on a finer scale. In the first example, shocks in fluids are discontinuities in fluid parameters that can occur, say, as the boundary of the wake of a projectile moving supersonically. While the shocks represent singularities in the fluid equations, they do not concern us in a fundamental way since the underlying particles composing the fluid do not individually have discontinuous or singular parameters. The second well-known singular behavior is that exhibited by a classical system consisting of a negative point particle orbiting an equal and oppositely charged massive central particle. This motion produces electromagnetic radiation that causes the orbiting particle to spiral into the central particle. This phenomenon became a crisis for classical electromagnetism when it was discovered that the distribution of mass and charge in atoms follows this model. The predicted collapse of the classical atom would make it impossible for matter to exist. Fortunately, as we now know, atoms described by the laws of quantum mechanics do not collapse, resolving this crisis.

Singularities also arise in the simplest solutions in general relativity. The static Schwarzschild solution [21.1, 2] describes the spacetime outside a spherically symmetric mass M. If the system under study is a black hole, in terms of the coordinate r that labels each two-sphere, as measured by an observer at $r = \infty$, various quantities such as a redshift behave badly at the event horizon, $r = 2GM/c^2$. This coordinate singularity may be removed by transforming to Kruskal coordinates [21.2, Section 6.4]. In Kruskal coordinates it is easy to see that an observer falling through the horizon would experience infinite tidal force as $r \to 0$ as measured by the Riemann tensor. This type of singularity

is called *curvature blow-up* since nonzero, coordinate-invariant quantities formed from the Riemann tensor become infinite as $r \rightarrow 0$. Because, inside the event horizon, r is a time-like variable, the singularity occurs on a space-like hypersurface.

The simplest cosmological solutions, the Friedmann–Robertson–Walker (FRW) cosmologies [21.1, 2], assume that the metric and matter variables depend only on time and that any observer at fixed spatial coordinates would measure the same values of all the variables such as density and expansion rate independent of spatial location or direction of observation. Such solutions are said to be spatially homogeneous and isotropic. When evolved backward in time, in the collapsing direction defined by the decrease of any fiducial volume with time, density, temperature, and tidal force become infinite after a finite time. This singularity, the big bang, creates a boundary to the evolution in the collapse direction and thus represents a past boundary to evolution in the expansion direction. In terms of classical general relativity, there is no way to ask what came before the big bang. The big bang is also a curvature blow-up singularity occurring on a space-like hypersurface.

Both the Schwarzschild and FRW solutions are physically relevant; the former to describe nonrotating black holes and the latter as a model for the universe on the largest scales. Both solutions have curvature blow-up singularities that occur on a space-like hypersurface. As we shall see in the following sections of this chapter, an important question is whether these simplest solutions are typical in the character of their singularities. Theorems developed originally by Penrose and Hawking [21.2] state that singular behavior should be expected in generic collapse once gravity becomes sufficiently strong. However, the theorems do not provide any details of the singularities one should expect in any particular case. In the following, we shall start with the FRW solutions and add increasing complexity by eliminating the restrictions to isotropy and spatial homogeneity. A combination of mathematical, heuristic, and numerical methods can be used to explore the behavior to be expected generically in these models. How worried one should be about the proliferation of singular behavior in general relativity is an area of active research. Unresolved questions include the detailed nature of generic singularities, whether or not singularities can form in our universe, and whether they have observational consequences. It is also not known if the appeal to an underlying quantum gravity theory will resolve singularities as quantum mechanics (or quantum electrodynamics) has done for electromagnetism. Promising results along these lines are discussed elsewhere in this volume.

21.1.2 FRW Models in the Collapse Direction

Discussions of cosmology are found elsewhere in this volume. The brief introduction given here, including the equations given in this subsection, may be found in most general relativity textbooks, such as [21.1]. Not too long after Einstein introduced the general theory of relativity, the first cosmological solutions were obtained. The goal of these solutions was to model the universe itself as characterized by stars and galaxies distributed everywhere and in all directions. At the time, the simplest assumptions to make were to look for solutions of Einstein's equations that depended only on time – to look for solutions that were identical at all spatial points and for all directions – spatially homogeneous and isotropic. Solutions of this type were quickly found but had two odd properties that made them seem unphysical. First, for reasonable matter such as pressureless particles (e.g., stars) called *dust*, no static solution was possible. This led Einstein to propose the cosmological constant to force a balance. He later abandoned the constant when it was discovered that the universe was indeed not static, but was expanding. Much more recently, the observed acceleration of the universe has led to the return of the cosmological constant in the guise of dark energy. The metric is given by

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + R(t)^2 \times \left(\frac{\mathrm{d}r^2}{1-kr^2} + r^2\,\mathrm{d}\theta^2 + r^2 \sin^2\theta\,\mathrm{d}\phi^2 \right) , \tag{21.3}$$

where t is comoving proper time, the time measured by observers at rest in the spatial coordinates, and r,θ, and ϕ are polar coordinates. $R(t)$, often called $a(t)$ in the literature, is the scale factor that measures the proper length of any spatial coordinate distance. The spatial topology is determined by the value of k which is -1, 0, or $+1$ for open, flat, and closed models, respectively. The matter in these models may be assumed for simplicity to satisfy a perfect fluid equation of state with energy density $\rho(t)$ and pressure $p(t)$. The Bianchi identities for the stress–energy tensor then yield, e.g.,

$$\rho_{\text{dust}} = \rho_0 \left(\frac{R_0}{R}\right)^3 ; \quad \rho_{\text{radiation}} = \rho_0 \left(\frac{R_0}{R}\right)^4 , \tag{21.4}$$

for dust ($p = 0$) and radiation ($p = \frac{1}{3}\rho$), respectively, where p is the pressure of the fluid. Given the equation of state of the matter, the dynamics of these models may be obtained from one of Einstein's equations, the energy-like Hubble's equation, namely

$$\left(\frac{\dot{R}}{R}\right)^2 + \frac{k}{R^2} - \frac{\Lambda}{3} = \frac{8\pi}{3}\rho\,, \tag{21.5}$$

where the constants G and c have been set equal to unity and Λ is the cosmological constant. The solutions have the form $R \propto t^{2/3}$ for dust and $R \propto t^{1/2}$ for radiation if, e.g., $k = 0 = \Lambda$.

The second odd property of these models was the *singularity* apparent in the limit as $t \to 0$ for t the time variable given in (21.3). In this limit, any fiduciary volume vanishes while the density becomes infinite a finite time in the past, namely, $R^3 \to 0$ as $t \to 0$. We note further from (21.4) and (21.5) that radiation dominates dust and both dominate over the curvature term k/R^2 and Λ as $t \to 0$. More formal definitions of singular behavior [21.2] demonstrate that the local tidal force as measured from invariants of the Riemann tensor also becomes singular. If we assume that FRW cosmologies describe our actual universe, the singular behavior occurs a finite time in the past and no solution exists before that time. The spacetime cannot be extended through the singularity within classical general relativity. Note that the singularity occurs everywhere in the space-like hypersurface described by (r, θ, ϕ) in (21.3).

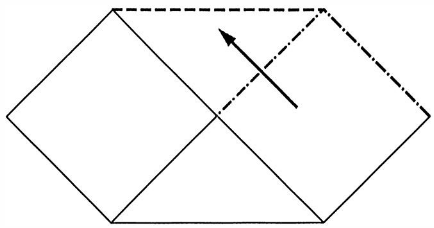


Fig. 21.3 The Penrose conformal diagram for the extended Schwarzschild spacetime is shown [21.2]. The *light grey arrow* represents an infalling light ray. The singularity at $r = 0$ is shown by the *brown dashed line*, the event horizon at $r = 2M$ by the *light brown line with long and short dashes*, and null infinity by the *brown dash-dot-dot line*. Note that the only way to detect the presence of such a space-like singularity is to fall into it

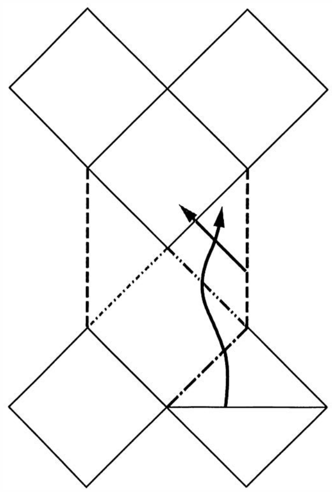


Fig. 21.4 The Penrose conformal diagram for the extended Reissner–Nordstrom spacetime [21.2]. The *horizontal line* represents space-like initial data. The singularity is time-like, represented by *vertical dashed lines* (*brown*). The outer horizon is the *brown line with long and short dashes*. The *light brown dash-dot line* is the inner horizon. The *gray small dash-dot line* is a Cauchy horizon, marking the boundary of causality of the initial data. The *light brown curve* represents a typical infalling observer that can detect the presence of the singularity via the (*light grey*) light ray emitted by it

21.1.3 Singularity Theorems

In the 1960s, Penrose and Hawking separately and jointly published papers proving that for reasonable matter starting from well-behaved initial conditions, if the gravitational field becomes sufficiently strong, the future evolution will yield some type of singular behavior. Precise statements and original references may be found in [21.2]. These theorems mean that singularities are a generic feature of general relativity. There are many ways to state the difficulties posed by such singularities. For example, singularities are a barrier to the indefinite evolution of spacetimes from well-posed initial data. Furthermore, the breakdown of general relativity at a singularity means that, were one to form in our universe, its influence on the surrounding spacetime would be unpredictable.

Of course, the well-known example of flat spacetime is nonsingular. However, it is not generic. While genericity of a spacetime is somewhat in the eye of the beholder, one usually requires the initial conditions to comprise an open set in a parameter space that is as large as possible and to lack any special

symmetry. A related consideration would be the effect of a singularity on our universe if one either formed or existed from the beginning. These concerns are addressed in Penrose's cosmic censorship conjectures (see, e.g., [21.2]). The cosmological application of the weak cosmic censorship conjecture states that singularities in our universe are only found as an unreachable past big bang. In the noncosmological, black hole scenario, weak cosmic censorship requires singularities to be hidden from the external world by a horizon. We could feel free to ignore big bang cosmological singularities and black hole singularities in practice, although not in principle.

Penrose introduced conformal diagrams to illustrate the properties of singularities. A conformal factor changes lengths and times in the spacetime metric but leaves the paths of light rays invariant. Thus these diagrams emphasize the paths of light rays in the spacetime. Examples of Penrose diagrams for the extended Schwarzschild and Reissner–Nordstrom spacetimes are given in Figs. 21.3 and 21.4, respectively (see also [21.2, Chap. 12]). In both figures, the light rays propagate at $\pm 45°$. For the extended Schwarzschild spacetime, the figure shows the space-like singularity at $r = 0$ and the null or light-like event horizon. Light rays cannot propagate forward in time from the singularity so a time-like or null observer will receive no information from the singularity before falling into it. In contrast, as shown in Fig. 21.4, the singularities in Reissner–Nordstrom (and Kerr) spacetimes are time-like. Light rays emitted by the singularity can impinge upon an infalling time-like observer.

The strong cosmic censorship conjecture would, if true, rule out time-like singularities that arise generically. Such a singularity could be detected by an observer falling into it. Problematical time-like singularities are found in the interiors of Kerr, Kerr–Newman, and Reissner–Nordstrom black holes. In contrast, the singularities in Schwarzchild black holes and in the FRW cosmology are space-like, in agreement with cosmic censorship. None of these examples address the question of cosmic censorship in generic models. Since the focus in this chapter is cosmological singularities, we will emphasize the question of whether or not generic singularities are space-like as the relevant aspect of cosmic censorship.

21.2 Spatially Homogeneous Cosmological Spacetimes

21.2.1 Introduction to Bianchi Type Spacetimes

If, in the spatial metric, we relax the assumption of isotropy, we find that there are many spatially homogenous cosmologies. While this may appear surprising, such spacetimes need only be solutions to Einstein's equations where the spatial variables do not appear in the Einstein equations although they may appear in the metric. An example of a two-dimensional homogeneous space with a spatially dependent metric in a coordinate basis is the two-sphere where the metric can be written as $dl^2 = d\theta^2 + \sin^2\theta \, d\phi^2$. It is clear that all points on the surface of the sphere are equivalent.

Three-dimensional, spatially homogeneous spaces were classified by Bianchi and include most of the interesting spatially homogeneous spaces. We are including a discussion of the classification for completeness. The details are not necessary for the exploration of particular homogeneous spaces. For precise definitions of Bianchi's classification, see [21.2]. For a detailed discussion of his classification and the resultant Bianchi types, see [21.3, 4]. The classification identifies the three Killing vectors ξ_i for $i = 1, 2, 3$ of the homogeneous space and is determined by the commutators (or Lie group) that relates them, namely

$$[\xi_i, \xi_j] = C^k_{ij} \xi_k \,, \tag{21.6}$$

where the properties of the *structure constants* C^k_{ij} determine the Bianchi type. One can also describe these spaces in terms of an invariant coordinate basis X_i such that $[X_i, X_j] = -C^k_{ij} X_k$ and the dual 1-forms σ^i such that

$$d\sigma^k = C^k_{ij} \sigma^i \wedge \sigma^j \,. \tag{21.7}$$

See, e.g., [21.5, p. 110], for more details. To construct a cosmological spacetime from any of the Bianchi type homogeneous spaces, scale the space with an overall time-dependent factor of the spatial volume and add the $t - t$ component of the metric. It should be emphasized here that the Bianchi type homogeneous spaces are constructed from the one-forms σ^i according to the prescription given by the structure constants in (21.7). A representation in a coordinate basis as, e.g., $\sigma^i = a^i_x(x,y,z)\, dx + a^i_y(x,y,z)\, dy + a^i_z(x,y,z)\, dz$ can

yield a spatially dependent metric. However, geometrical quantities appearing in Einstein's equations such as the Ricci tensor and the spatial scalar curvature will depend only on time.

Under some circumstances, it might be convenient to add a lapse N and/or shift N_i to allow a more general time variable in the metric as written in a coordinate basis

$$\mathrm{d}s^2 = -N^2\,\mathrm{d}t^2 + 2N_i\,\mathrm{d}t\,\mathrm{d}x^i + g_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j\,. \quad (21.8)$$

Several of the cosmological spacetimes constructed from Bianchi type spaces are useful for the study of cosmological singularities. In this context, it is convenient to describe the homogeneous spaces through a spatial metric that will emphasize the anisotropy. The components of the spatial metric will be written as functions of time multiplying the one-forms σ^i. As an example, the flat FRW model's spatial metric

$$\mathrm{d}l^2_{\mathrm{FRW}} = R^2(t)(\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2)\,, \quad (21.9)$$

will become for an anisotropically expanding, flat model

$$\mathrm{d}l^2 = R_x^2(t)\,\mathrm{d}x^2 + R_y^2(t)\,\mathrm{d}y^2 + R_z^2(t)\,\mathrm{d}z^2\,. \quad (21.10)$$

To study the anisotropic cosmologies, it is convenient to replace the anisotropic scale factors $\{R_x, R_y, R_z\}$ with $\{\Omega, \beta_+, \beta_-\}$, where, in this example, $\Omega = \ln(R_xR_yR_z)^{1/3}$ measures a fiducial spatial volume while $\beta_+ = \frac{1}{6}\ln(R_yR_z/R_x^2)$ and $\beta_- = \frac{1}{2\sqrt{3}}\ln(R_yR_z)$ measure the anisotropic shear. For the cosmological spacetimes discussed below, these variables allow the evolution of the cosmology to be treated as a trajectory in an abstract space called minisuperspace (MSS) with axes $\{\Omega, \beta_+, \beta_-\}$ [21.5].

MSS was introduced originally by *Misner* [21.6]. Superspace in this context (there are other completely different uses of this term) is an abstract space wherein each point represents a three-geometry. A three-geometry generalizes the concept of a space to take into account that many apparently different spaces are actually the same space written in different coordinates. A three-geometry is the equivalence class of such spaces – the coordinate invariant underlying three-dimensional space. A trajectory in superspace may be interpreted as the evolution of a space in time. For example, the FRW cosmology may be represented in superspace by a single parameter, e.g., the spatial volume. A straight line trajectory could then represent the evolution of the volume with time. In principle, however, superspace is infinite dimensional where an arbitrary space may be specified (modulo coordinate transformations) by the set of numbers representing, e.g., the metric values at each spatial point. MSS is the restriction of these infinite degrees of freedom to the finite degrees of freedom (e.g., the metric coefficients valid everywhere) of a spatially homogeneous cosmology. It is useful to analyze spatially homogeneous cosmologies through their dynamics in MSS, especially through Hamiltonian methods using the dynamical variables $\{\Omega, \beta_+, \beta_-\}$ and their conjugate momenta $\{p_\Omega, p_+, p_-\}$.

Without the restriction to flat space, the spatial metric is given by

$$\mathrm{d}l^2 = \left(\mathrm{e}^{2\beta}\right)_{ij}\sigma^i\sigma^j\,, \quad (21.11)$$

where β_{ij} is called the anisotropy matrix and σ^i is the basis of one-forms in (21.7). The anisotropy matrix is constructed to be traceless since the trace appears in the determinant of the spatial metric (21.11), which is, in fact, the volume density. A cosmology is then constructed by adding an overall isotropic scale factor $\mathrm{e}^{2\Omega}$ where Ω is proportional to the logarithm of the spatial volume. It should be noted that the sign of Ω varies in the literature. Thus, we obtain the metric

$$\mathrm{d}s^2 = -N^2\,\mathrm{d}\tau^2 + \mathrm{e}^{2\Omega}(\mathrm{e}^{2\beta})_{ij}\sigma^i\sigma^j\,, \quad (21.12)$$

where we have called the coordinate time τ and added an arbitrary lapse $N(\tau)$. For certain of these models, called class A [21.4], Einstein's equations may be obtained by variation of a Hamiltonian

$$2\mathcal{H} = -p_\Omega^2 + p_+^2 + p_-^2 + \mathrm{e}^{4\Omega}V(\beta_+, \beta_-)\,, \quad (21.13)$$

where $p_\Omega, p_\pm$ are canonically conjugate to $\Omega, \beta_\pm$, the potential $V(\beta_+, \beta_-)$ arises from the spatial scalar curvature, and $\mathcal{H} = 0$ is the Hamiltonian constraint for these models. The equations obtained by the variation of the Hamiltonian (21.13) can be interpreted to describe a trajectory in MSS that we shall see can be treated in detail as a scattering off the MSS potential $\mathrm{e}^{4\Omega}V(\beta_+, \beta_-)$. We note here that Einstein's equations can be formulated in terms of first-order-in-time evolution equations and constraints. The constraint equations, once solved initially, are preserved for all time by the evolution equations. In the spatially homogeneous cosmologies discussed in this chapter, the Hamiltonian

constraint also provides the Hamiltonian whose variation yields the evolution equations. It should be noted that the preservation of the constraints by the evolution equations is not guaranteed in a numerical evolution.

While there are other spatially homogeneous spacetimes that may be interpretable as cosmologies, most features of generic singularities may be captured through only three Bianchi types – I, II, and IX. Each of these will be described in detail in subsequent sections.

21.2.2 Matter (Usually) Does Not Matter

To study the approach to the singularity, consideration of the vacuum case is usually sufficient. This is definitely not true for FRW models where no solution to Einstein's equations exists in the absence of matter (or of a cosmological constant in place of matter). However, as we show in the following section, anisotropy can substitute for matter to construct a solution. The key to understanding the influence of matter on the dynamics of the approach to the singularity is the power-law dependence of the matter's energy density on the scale factor $R(t)$ as discussed above, where the fastest growth in the density is proportional to R^{-4} for radiation. We shall see below that an effective energy density due to anisotropy behaves as R^{-6} in the same variables. Thus, anisotropy will dominate as the scale factor goes to zero so that dust, radiation, or any other power law dependence that grows more slowly than anisotropy can be neglected. While the dust and radiation equations of state describe approximate behaviors of known matter, it is also possible that scalar fields play a role in the early universe. These approach the singularity with the same power law dependence on the scale factor as anisotropy energy density and thus cannot be neglected. In addition, neglecting matter – studying singularities in vacuum spacetimes – allows avoidance of nongravitational singularities such as shocks in fluids that can arise in the matter even in flat spacetime.

21.2.3 Examples

In this section, three examples of spatially homogeneous, anisotropic, vacuum cosmologies will be considered. We will treat these models as dynamical systems described as *particle* trajectories in MSS. Here anisotropy contributes the kinetic energy and spatial scalar curvature (present in all Bianchi types except I) contributes the potential energy. Hamiltonian methods can be used to derive Einstein's equations to yield the dynamics of a free particle that bounces off potential energy walls. While the three examples we will discuss allow this type of analysis, in general, Hamiltonian methods that assume symmetries such as spatial homogeneity prior to variation do not necessarily yield the correct Einstein's equations. Bianchi-type models in class B [21.4] require this more general treatment and will not be discussed here.

The Kasner Spacetime

The simplest vacuum, anisotropic, spatially homogeneous cosmological spacetime is the solution to Einstein's equations first described by *Kasner* [21.7], also known as Bianchi type I. The spatial coordinates may be either Cartesian or the equivalent angular variables on T^3. Rather than the single, isotropic FRW scale factor, the Kasner solution has three orthogonal scale factors R_x, R_y, and R_z. The metric is given by

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + \sum_{i=1}^{3} t^{2a_i}\,\mathrm{d}x_i^2\,, \tag{21.14}$$

where the exponents a_i satisfy two relations, namely

$$\sum_{i=1}^{3} a_i = 1 = \sum_{i=1}^{3} a_i^2\,, \tag{21.15}$$

and t is comoving proper time. Equations (21.15) imply that the three exponents may be replaced by a single parameter, u, where [21.8]

$$a_1 = \frac{-u}{u^2+u+1}\,;\quad a_2 = \frac{u+1}{u^2+u+1}\,;\quad a_3 = \frac{u(u+1)}{u^2+u+1} \tag{21.16}$$

for $1 \le u < \infty$. It is clear from (21.16) that, in the direction of a collapsing universe, two scale factors will be decreasing and one increasing. Note, however, that the overall spatial volume is proportional to $t^{a_1+a_2+a_3} = t$ and thus goes to zero at the big bang singularity in these models. The special case $u = \infty$ yields $a_1 = a_2 = 0$, $a_3 = 1$. With these values of the Kasner indices, the metric (21.14) is really flat spacetime in different coordinates. To see this, substitute (T, X) for (t, x_3) in (21.14), where

$$T = t\cosh x_3\,;\quad X = t\sinh x_3\,. \tag{21.17}$$

Part C | 21.2

To interpret the dynamics of anisotropic cosmologies, it is convenient to use the anisotropy and volume variables given in (21.12). In particular, we replace the Kasner metric (21.14) by

$$\mathrm{d}s^2 = -\mathrm{e}^{6\Omega}\,\mathrm{d}\tau^2 + \mathrm{e}^{2\Omega}(\mathrm{e}^{2\beta})_{ij}\,\mathrm{d}x^i\,\mathrm{d}x^j\,, \qquad (21.18)$$

where we have chosen $N = \mathrm{e}^{3\Omega}$ as the lapse while the anisotropy is parametrized by the diagonal matrix

$$\beta_{ij} = \mathrm{diag}\left(-2\beta_+, \beta_+ + \sqrt{3}\beta_-, \beta_+ - \sqrt{3}\beta_-\right)\,, \qquad (21.19)$$

and the time τ is defined by the choice of lapse. The set of variables $(\Omega, \beta_+, \beta_-)$ define the axes in MSS. Within MSS then, these variables are useful because Einstein's equations can be obtained by variation of the Hamiltonian

$$2\mathcal{H} = -p_\Omega^2 + p_+^2 + p_-^2\,, \qquad (21.20)$$

where p_Ω and $p_\pm$ are canonically conjugate to Ω and $\beta_\pm$, respectively. The resultant Hamiltonian (21.20) is equivalent to that for a free, relativistic particle in a three-dimensional MSS with axes $(\Omega, \beta_+, \beta_-)$. For this model, $\mathcal{H} = 0$ is, in fact, the Hamiltonian constraint of general relativity (see, e.g., [21.2], Section 10.2 and elsewhere in this volume) and is one of the Einstein equations. Of course, it is important to keep in mind that MSS is an abstract space so a free particle within it is an abstraction of the dynamics of a physical universe.

In terms of $R = \mathrm{e}^{\Omega}$, $\mathrm{d}t = \mathrm{e}^{3\Omega}\,\mathrm{d}\tau$ and (from variation of $\mathcal{H}$) $\mathrm{d}\Omega/\mathrm{d}\tau = -p_\Omega$. The Hubble equation (21.5) may be rewritten as

$$\left(\frac{\mathrm{d}\Omega}{\mathrm{d}t}\right)^2 = (p_+^2 + p_-^2)\,\mathrm{e}^{-6\Omega}\,, \qquad (21.21)$$

where the left-hand side is just $R^{-1}\,\mathrm{d}R/\,\mathrm{d}t$ and $p_\pm$ are constants. The right-hand side of (21.21) is an energy-like source for the evolution of the volume $\mathrm{e}^{3\Omega}$. It is clear that anisotropy can replace matter in these vacuum models and that the kinetic energy of the anisotropy can dominate ordinary matter as $t \to 0$, since radiation evolves as $\mathrm{e}^{-4\Omega}$ and dust as $\mathrm{e}^{-3\Omega}$ as $\Omega \to -\infty$.

Note that variation of the Hamiltonian (21.20) yields the equations of motion for a relativistic particle in an MSS of two *spatial* dimensions. It is convenient to solve the Hamiltonian constraint $\mathcal{H} = 0$ by defining new parameters $v_\pm = p_\pm/p_\Omega$ to yield

$$v_+^2 + v_-^2 = 1\,, \qquad (21.22)$$

which is immediately parametrizable in terms of θ as $v_+ = \cos\theta$ and $v_- = \sin\theta$ (since β_+ and β_- are taken as the horizontal and vertical axes, perpendicular to the Ω-axis in MSS). Note also that (21.22) defines a unit-radius circle in the space with axes $p_\pm/p_\Omega$. Any Kasner solution resides at a point on this *Kasner* circle. One can rewrite θ (nonuniquely) in terms of the previously mentioned parameter u [21.9]. Equations (21.20) and (21.22) may be combined to show that the Kasner solution may be represented by an arbitrary straight line in MSS, namely

$$\beta_\pm = \beta_\pm^0 - v_\pm\Omega\,, \qquad (21.23)$$

for $\beta_\pm^0$ constants of integration. As the cosmology evolves, the anisotropy and volume of the model evolve along the straight line.

It is also possible to calculate curvature invariants built out of the Riemann tensor for this model (see, e.g., [21.1]). The first nonzero invariant is

$$\zeta = R^{\mu\nu\rho\sigma}R_{\mu\nu\rho\sigma} = \frac{16}{t^4}\frac{u^2(u+1)^2}{(u^2+u+1)^3}\,, \qquad (21.24)$$

that is seen to blow up as comoving proper time $t \to 0$ for any finite value of u. This defines a curvature blow-up singularity in this model. Note that the singularity occurs at a fixed value of t (or Ω) and is thus space-like in character. The special case $u = \infty$, flat spacetime in disguise, is, as expected, nonsingular, since the coefficient of $16/t^4$ vanishes.

The approach to the singularity in Kasner (Bianchi type I) models is called velocity term dominated (VTD) [21.10] because the dynamics is described by the kinetic energy (velocity terms) of a free particle in MSS.

The Bianchi Type II (Taub) Models

The metric of the *Taub* [21.11] or Bianchi type II spacetimes may be written as (21.18) with $\sigma^i\sigma^j$ replacing $\mathrm{d}x^i\,\mathrm{d}x^j$ for $\mathrm{d}\sigma^1 = \sigma^2\wedge\sigma^3$, $\mathrm{d}\sigma^2 = 0$, $\mathrm{d}\sigma^3 = 0$. The spatial dependence necessary to enforce the group structure of the σ^i disappears from Einstein's equations (indicating that the model is, indeed, spatially homogeneous) that are derivable from the Hamiltonian

$$2\mathcal{H} = -p_\Omega^2 + p_+^2 + p_-^2 + \gamma^2\mathrm{e}^{4\Omega - 8\beta_+}\,, \qquad (21.25)$$

for γ an arbitrary constant, which we will set equal to unity for convenience. The collapsing direction is indicated by $\Omega \to -\infty$ and $\mathcal{H} = 0$ is the Hamiltonian constraint. In the direction toward the singularity, we may assume that $\Omega < 0$ and that the potential term in (21.25) may be written as

$$V = e^{4|\Omega|(1+2\beta_+/|\Omega|)} . \quad (21.26)$$

The exponential V attains its largest value in the allowed region

$$\frac{\beta_+}{|\Omega|} \geq -\frac{1}{2} , \quad (21.27)$$

when the equality holds. If the equality does not hold, $V \approx e^{-4|\Omega|\xi}$, where $\xi > 0$ and is thus exponentially small. As $|\Omega| \to \infty$, the difference between V at $\xi = 0$ and at $\xi > 0$ is amplified. Thus V becomes an ever sharper wall at $\frac{\beta_+}{|\Omega|} = -\frac{1}{2}$.

If the potential term is vanishingly small, the dynamics reverts to the Kasner solution. Because the potential term is exponential, it becomes ever smaller except where the exponent ≈ 0 as $\Omega \to -\infty$. Where the potential is exponentially small, the Kasner solution becomes an excellent approximation. Equation (21.25) describes the standard classical mechanics problem of scattering off an exponential potential, so that an explicit solution may be obtained. More important for the purposes of analyzing the approach to the singularity is to use conservation of momentum far from the potential where the potential can be neglected to relate the outgoing Kasner model to the incoming one. The key element of the calculation in this case is to rewrite (21.25) in terms of the variable $z = \Omega - 2\beta_+$ appearing in the exponential and an orthogonal (in a sense to be described) variable $y = a\Omega + c\beta_+$, where a and c are determined so that y, z, and their canonically conjugate momenta p_y, p_z satisfy

$$p_y\dot{y} + p_z\dot{z} = p_\Omega\dot{\Omega} + p_+\dot{\beta}_+ , \quad (21.28)$$

where the overdot denotes $\mathrm{d}/\mathrm{d}\tau$. This allows one to write (p_y, p_z) in terms of (p_+, p_Ω). In terms of these variables, the potential appears only in the z-direction so that the bounce law is obvious, namely p_z must change sign while p_y (and also, obviously, p_-) remain constant. One then rewrites the Hamiltonian (21.25) in terms of (y, z) and their conjugate momenta. To take the form of an energy equation for scattering off the potential e^{4z} requires $c = -a/2$ in the definition of y. One then rewrites (p'_+, p'_Ω) after the bounce in terms of the post-bounce $(p'_y, p'_z) = [p_y(p_+, p_\Omega), -p_z(p_+, p_\Omega)]$ to yield the bounce law (in terms of v_+ for convenience)

$$v'_+ = -\frac{5v_+ + 4}{4v_+ + 5} . \quad (21.29)$$

(The parameter a does not contribute to the bounce law and thus need not be determined.) Finally, we note that a rotation of the $\beta_\pm$ axes in MSS is equivalent to orienting the potential wall for this model in an arbitrary direction. This allows the conclusion that the dynamics of the Taub model is described as an incoming Kasner bouncing off an exponential wall in MSS to yield an outgoing Kasner with the change in Kasner parameter obtained from the bounce law (21.29) for v_+ and the constancy of v_- with $v_\pm$ rotated appropriately according to the orientation of the Taub potential wall in MSS. Note that this behavior can also be described as a trajectory from one point on the Kasner circle to another. The Taub approach to the (Kasner-like) singularity is called asymptotically velocity term dominated (AVTD) because, after the bounce, the dynamics becomes arbitrarily close to the VTD Kasner solution.

The Bianchi Type IX (Mixmaster) Models

In the context of singularity behavior, one might argue that the vacuum, diagonal, Bianchi type IX model, first described by Belinski, Khalatnikov, Lifshitz (BKL) [21.8] and dubbed *Mixmaster* by *Misner* [21.12], holds the most interest. We note here, but will not discuss further, that similar behavior exists for corresponding Bianchi type VIII models and for Bianchi type VI_0 models with magnetic fields. A review is given in [21.13]. The Mixmaster metric is given by (21.12) with anisotropy matrix (21.19), where the σ^i satisfy

$$\mathrm{d}\sigma^i = \epsilon^i_{jk}\sigma^j \wedge \sigma^k , \quad (21.30)$$

and where ϵ^i_{jk} is zero if any two indices are identical, $+1$ if they are in cyclic order, and -1 if they are anticyclic. Therefore, the symmetry group generated by the Killing vectors ξ^i is $SO(3)$. Realization of this model requires the spatial topology S^3. In an appropriate coordinate basis, the dynamics can be expressed in terms of three *logarithmic scale factors* (LSFs) that correspond to the logarithms of the three scale factors, $\ln R_x$, $\ln R_y$, $\ln R_z$, in the Kasner model. For Bianchi type IX, the LSFs are

$$\alpha = \Omega - 2\beta_+ ,$$
$$\zeta = \Omega + \beta_+ + \sqrt{3}\beta_- ,$$
$$\gamma = \Omega + \beta_+ - \sqrt{3}\beta_- . \quad (21.31)$$

In any Kasner epoch between bounces, two LSFs decrease and one increases in the direction toward the singularity. In subsequent epochs, the most negative LSF decreases monotonically while the other two exchange roles after each bounce. When the era ends as indicated by the BKL map given below for u from (21.16), one of the oscillating LSFs becomes the monotonically decreasing one while the other two now oscillate. The LSFs are proportional to the logarithms of the dominant potential terms described in the Hamiltonian treatment given next.

The detailed dynamics may be found from variation of the Hamiltonian (21.13) where the potential V is given by

$$\begin{aligned} V(\beta_+, \beta_-) &= \mathrm{e}^{-8\beta_+} + \mathrm{e}^{4\beta_+ + 4\sqrt{3}\beta_-} + \mathrm{e}^{4\beta_+ - 4\sqrt{3}\beta_-} \\ &\quad - 2\left(\mathrm{e}^{4\beta_+} + \mathrm{e}^{-2\beta_+ - 2\sqrt{3}\beta_-} + \mathrm{e}^{-2\beta_+ + 2\sqrt{3}\beta_-}\right). \end{aligned} \tag{21.32}$$

Under most circumstances, the first three terms dominate. Note that these terms are equivalent to three copies of the Taub potential rotated with respect to each other by $\frac{2}{3}\pi$. The subdominant terms become significant only in the corners of the potential shown in Fig. 21.5. The approximate behavior of this model is a sequence of Kasner epochs related by the so-called BKL map [21.8] except when, in MSS, the system point runs into one of the *corners* of the potential. This yields a different dynamical picture as described in, e.g., [21.14]. Eventually, the system will exit the corner and patch onto the generic, sequential Kasner dynamics. If the trajectory bounces from the center of an exponential wall precisely down the center of the corner (e.g., a bounce at $\beta_- = 0$ with $p_- = 0$), no further bounces will occur. This not only corresponds to the Taub solution but the final Kasner is the exceptional flat spacetime case, $u = \infty$. To analyze the behavior of Mixmaster dynamics, it is useful to follow the behavior of the BKL parameter u defined for each Kasner segment, or epoch, via (21.16). The successive values of u obey a discrete map, the BKL map, describing the evolution of u, between Kasner *epochs* labeled by n and $n+1$

$$u_{n+1} = \begin{cases} u_n - 1 & u_n \geq 2 \\ \dfrac{1}{u_n - 1} & 1 \leq u_n \leq 2 \end{cases} \tag{21.33}$$

A detailed treatment of an updated version of the four-parameter BKL map is given in [21.15]. It is shown there that one can construct an approximate Mixmaster trajectory from the map and, in [21.16], that the approximation improves as the singularity is approached. See Fig. 21.6 and the discussion in [21.16]. The map for u has been analyzed in great detail by several authors [21.17]. Each Kasner epoch is related to the previous one by $u_{n+1} = u_n - 1$. This set of epochs constitutes an *era* as long as $u_n > 2$. The eras themselves are related by the so-called Gauss map $u_{N+1} = 1/(u_N - [u_N])$, where $[u_N]$ is the integer part of u_N. Here, N labels the eras rather than the epochs within an era. The Gauss map is sensitive to initial conditions due to the subtraction in the denominator. This is one of the characteristics of chaotic dynamics. For several years, there was extensive discussion of the relationship of this property to chaos and its various definitions. The conclusion one may draw is that this is really just a semantic question with the answer that the Gauss map is sensitive to initial conditions whether or not this is labeled *chaos*. The Gauss map has other interesting properties, including a fixed point at the golden mean ($u = (1+\sqrt{5})/2$) and period-n solutions that relate to the Fibonacci series [21.9]. This countable set within the continuum of u-values complicates proofs of the

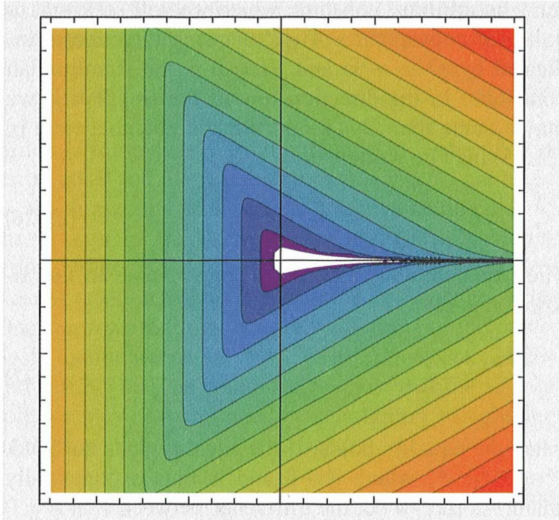


Fig. 21.5 The contours of the Mixmaster potential from (21.32) are shown using a logarithmic scale, $\log V(\beta_+, \beta_-)$, where the *horizontal axis* is β_+ and the *vertical* β_- with the range $[-5, 5]$. The potential is negative in the region appearing *white* in the figure. The straight *walls* of the equipotentials meet at *corners*. As $\Omega \to -\infty$, the Mixmaster trajectory reaches ever larger $\beta_\pm$ so that the walls become steeper and the corners narrower for successive epochs

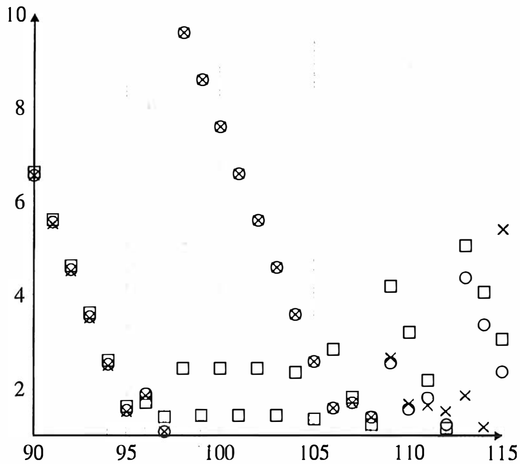


Fig. 21.6 Accuracy of the BKL map for u. The map (21.33) is calculated to 60 significant figures using *Mathematica* (shown by *crosses*) and compared to u obtained from numerical simulations with double precision (*squares*) and quadruple precision (*circles*) using the algorithm of [21.16]. The *horizontal axis* shows the epoch number n and the *vertical axis* the value of u. While all these calculations agree initially, sensitivity to initial conditions eventually causes disagreement. The sequences of u-values from the simulations begin to disagree with the highest precision calculation when the information about the initial value of u is lost. The degree of initial information depends on the precision with which it is provided. The double precision simulation loses its initial precision at $n = 98$ and the quadruple at $n = 113$. Once the initial information is *lost* the subsequent evolutions differ qualitatively, where qualitative difference means that the sequences of integer parts of u begin to differ

nature of the dynamics. Especially problematic are precisely integer values of u, which lead exactly to the Taub solution's special case of the flat spacetime Kasner mentioned above.

Numerical methods have been brought to bear to solve the ordinary differential equations for this model from its earliest history. Typically, only a few Kasner epochs have been shown due to loss of accuracy and the need to reduce the computational time step during a bounce. The best numerical method now in use was developed by *Garfinkle* et al. [21.16] and is based on a symplectic algorithm related to that first applied to cosmological singularities in [21.18]. A brief discussion of symplectic numerical methods will be given next.

Following the discussion in [21.19], we review symplectic numerical methods for a system with one degree of freedom q with conjugate momentum p described by a Hamiltonian

$$H = \frac{1}{2}p^2 + V(q)\,. \tag{21.34}$$

To illustrate the method, we choose

$$H = H_1(p) + H_2(q)\,. \tag{21.35}$$

If the vector $X = (p, q)$ defines the variables at time t, then the time evolution is given by

$$\frac{\mathrm{d}X}{\mathrm{d}t} = \{H, X\}_{\mathrm{PB}} \equiv AX\,, \tag{21.36}$$

where $\{\}_{\mathrm{PB}}$ is the Poisson bracket. The usual exponentiation yields an evolution operator

$$\mathrm{e}^{A\Delta t} = \mathrm{e}^{A_1(\Delta t/2)}\mathrm{e}^{A_2\Delta t}\mathrm{e}^{A_1(\Delta t/2)} + O(\Delta t^3)\,, \tag{21.37}$$

where $A = A_1 + A_2$ is the generator of the time evolution. This method is useful when exact solutions for the sub-Hamiltonians are known. For the split given in this example, variation of H_1 yields the solution

$$q = q_0 + p_0\Delta t\,, \quad p = p_0\,, \tag{21.38}$$

while that of H_2 yields

$$q = q_0\,, \quad p = p_0 - \left.\frac{\mathrm{d}V}{\mathrm{d}q}\right|_{q_0} \Delta t\,. \tag{21.39}$$

Note that H_2 is exactly solvable for any potential V no matter how complicated. One evolves from t to $t + \Delta t$ using the exact solutions to the sub-Hamiltonians according to the prescription given by the approximate evolution operator (21.37).

The optimal application to Mixmaster models requires the Hamiltonian to be split as $H_1 = H_{\mathrm{taub}}$; $H_2 = H_{\mathrm{other}}$ where H_{taub} is (21.13) using the largest exponential term in (21.32) for V, while H_{other} contains the remaining terms. Evolution with H_1 uses the exact Taub solution while the evolution with H_2 is straightforward since no momenta are involved. Since H_2 is exponentially small in this case, the accuracy of the method is exponential rather than that of the nominal order of the symplectic scheme (Fig. 21.7). Essentially, one evolves the system as a sequence of Taub models. This method allows evolution through hundreds

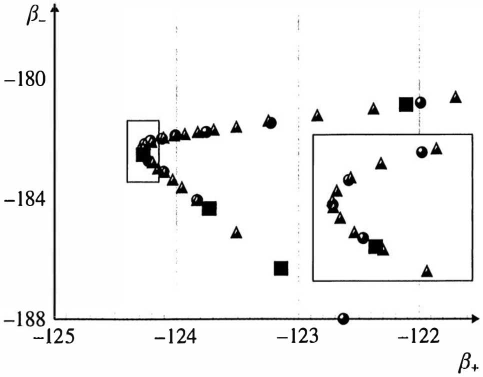


Fig. 21.7 Performance of the algorithm described in [21.16] (where a version of this figure appears). A typical Mixmaster bounce is shown. First *triangles* and *circles* indicate every tenth point in, respectively, fourth-order Runge–Kutta and sixth-order symplectic (split into kinetic and potential sub-Hamiltonians) integration schemes. In contrast, the *solid squares* show *every* point using the new algorithm. The *inset* shows the bounce in more detail. Because the bounces are built in, the new algorithm can continue to increase the time step to match the increase in magnitude of Ω, $\beta_\pm$. In standard methods, the time step must be decreased to follow the details of the bounce

of epochs with ever-increasing accuracy. The accuracy improves because the neglected terms become exponentially smaller as the singularity is approached. A typical trajectory obtained with this algorithm and using rescaled coordinates $\beta_\pm/|\Omega|$ is shown in Fig. 21.8. It shows the ergodic nature of the dynamics in that the trajectory appears to reach every part of the available region in the rescaled MSS.

It should be noted at this point that attempts to exploit the Mixmaster behavior to explain features of our universe have failed. This is because any fiducial time indicator such as the ratio ψ of the Hubble time to the Planck time fails to allow more than a few bounces, not enough to, e.g., isotropize the universe or wash out any special initial conditions [21.12]. The change in the relevant time variable, e.g., $\Delta\Omega$ in the approach to the singularity, is the logarithm of ψ and reaches at most of order 10^3, while on the order of 250 bounces to allow loss of information about initial conditions corresponds to $\Delta\Omega \approx 10^{60}$.

It is also worth mentioning that the Mixmaster bounces end during the evolution toward the singularity in the presence of a scalar field [21.20, 21]. Unlike dust or radiation, the kinetic energy of the scalar field

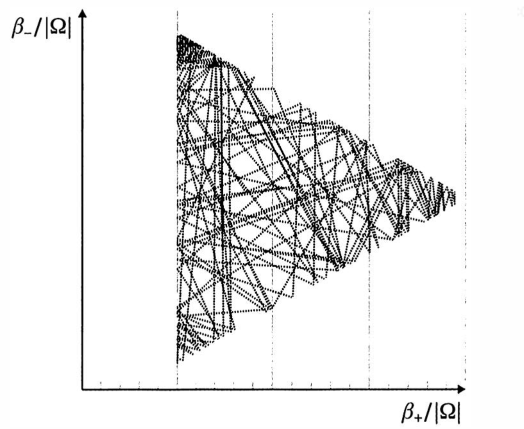


Fig. 21.8 A typical Mixmaster trajectory with ≈ 250 bounces. In these rescaled variables, the potential walls are fixed with, e.g., the wall due to the term $e^{4\Omega-8\beta_+}$ fixed at $\beta_+/|\Omega| = -\frac{1}{2}$

enters the Hamiltonian constraint with the same power-law dependence on the volume as the anisotropy energy to yield a Hamiltonian

$$2\mathcal{H} = -p_\Omega^2 + p_+^2 + p_-^2 + p_\varphi^2 + e^{4\Omega} V(\beta_+, \beta_-) + e^{6\Omega} m^2 \varphi^2 \,, \tag{21.40}$$

where φ, p_φ are, respectively, a minimally coupled scalar field of mass m and its canonically conjugate momentum and $\mathcal{H} = 0$ is the Hamiltonian constraint. Equation (21.40) implies that the scalar field momentum can influence the dynamics of the approach to the singularity because the term p_φ^2 is not suppressed by factors of e^Ω as $\Omega \to -\infty$. What happens is that the Kasner-like terms in the kinetic energy in (21.40) that dominate between bounces no longer satisfy the Hamiltonian constraint by themselves. The kinetic energy is also shared by the scalar field. The system point in MSS with coordinates $(\Omega, \beta_+, \beta_-)$ no longer moves fast enough to hit the potential wall so the bounces come to an end. Thus Bianchi type IX models with minimally coupled scalar fields are AVTD. An example is shown in Fig. 21.9 and additional details can be found in [21.21].

It is important to mention here that the Hamiltonian constraint $\mathcal{H} = 0$ must be preserved during the simulation in models with Mixmaster-like behavior (including the spatially inhomogeneous models to be discussed later). See [21.22] for more details. For example, if

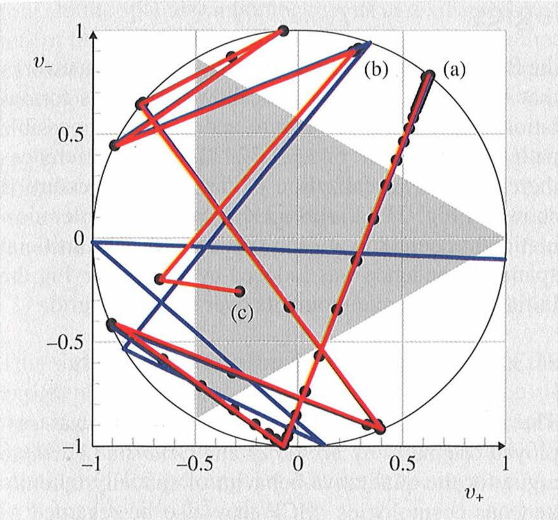


Fig. 21.9 Trajectory of a Mixmaster model with a scalar field ends. The Kasher circle defined by $v_+^2 + v_-^2 = 1$ is shown. Two Mixmaster trajectories are shown, one with (*red with dots*) and one without (*blue*) a scalar field. Both start near (*a*). The nonscalar-field trajectory returns to the Kasner circle after each bounce. The scalar-field trajectory has $v_+^2 + v_-^2 = 1 - v_\varphi^2$ and fails to reach the Kasner circle after a few bounces. The scalar-field trajectory begins to deviate from the nonscalar-field one at (*b*) and ends at the point labeled *c*. A version of this figure appears in [21.21]

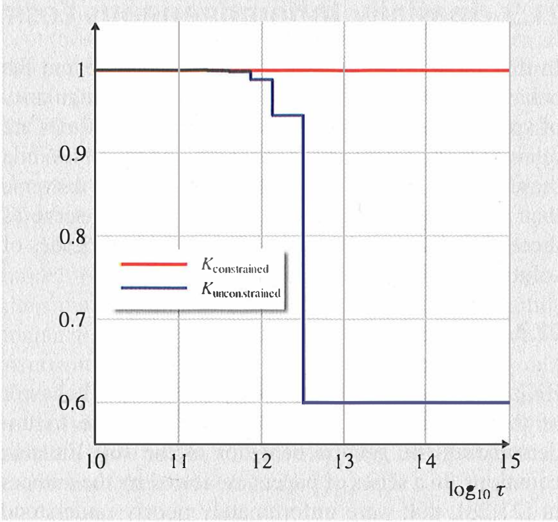


Fig. 21.10 Preserving the constraint in Mixmaster simulations. The value of $K = v_+^2 + v_-^2$ versus computational time τ is shown for a typical simulation with enforcement of the Hamiltonian constraint $\mathcal{H} = 0$ (*red*) and nonenforcement (*blue*). Between bounces, the true solution should have $K = 1$. This indicates that failure to preserve the constraint during numerical evolutions can yield spurious results

we consider Mixmaster evolution using (21.13) with (21.32) for $\mathcal{H}$, the constraint can be enforced by solving $\mathcal{H} = 0$ for p_Ω at every N-th time step rather than using p_Ω obtained via the numerical evolution algorithm. Results from the method of [21.16] are not very sensitive to the value chosen for N, although the precision and information loss shown in Fig. 21.6 will depend on N after a large number of bounces. As $\Omega \to -\infty$, the bounces become ever sharper so that the system spends most of the time on the Kasner circle. Thus the value of $K = v_+^2 + v_-^2$ on the Kasner circle (defined by $K = 1$) may be used to test the validity of the simulation. This is shown in Fig. 21.10.

For spatially homogeneous models, the mathematical status of Mixmaster dynamics is reasonably complete. The simpler cases with VTD or AVTD singularities such as Bianchi types I and II present no difficulties. However, Bianchi types VIII and IX with Mixmaster behavior have proven to be much more challenging. In this context, *Ringström* has proven that, if the initial data are not Taub, the diagonal, Bianchi VIII and IX, vacuum models have curvature blow-up singularities [21.23]. A similar proof was given by *Weaver* for the approach to the Mixmaster-like singularity in magnetic Bianchi VI_0 [21.24]. *Wainwright* and collaborators [21.25] developed a framework to unify the Bianchi types. The spirit of this framework is an alternative view of the Bianchi type models as dynamical systems with the structure focusing on the Kasner circle defined by (21.22) rather than as dynamics within the MSS potential. With these variables and extensions thereof, it has been possible to prove [21.26, 27] that the behavior described above – the unending sequence of Kasner epochs related to each other by the BKL map – characterizes generic Bianchi type IX models as the singularity is approached.

21.3 Spatially Inhomogeneous Cosmologies

In this section, we shall determine to what extent the behaviors in the approach to the space-like singularity of spatially homogeneous cosmological spacetimes are relevant in the presence of spatial inhomogeneity. While these models will be no closer to the actual universe than those in the previous section, they will serve as theoretical laboratories to explore generic properties of solutions to Einstein's equations.

21.3.1 The BKL Conjecture

BKL's primary motivation to understand the behavior of the Mixmaster universe was as a prelude to understanding the generic behavior of the full Einstein equations. In a series of papers, reviewed by the authors in [21.28], that were unfortunately poorly understood by others, they studied the evolution in the collapsing direction of a spatially inhomogeneous cosmology with variation in only one spatial direction and time. They determined that such a model was unstable to variations in the other spatial directions and concluded that one would eventually obtain an evolution behaving as a separate spatially homogeneous Mixmaster model at every spatial point. This conclusion requires the eventual dynamics to be dominated at every spatial point by time derivatives (i. e., kinetic energy-like terms) rather than spatial gradients. In special cases, the asymptotic behavior could be Kasner-like rather than Mixmaster-like at every spatial point. A simple way to visualize the BKL conjecture is that, after some time (which may be different at different spatial points), the partial differential equations of general relativity may be replaced by separate ordinary differential equations at every spatial point and that the ordinary differential equations are those for spatially homogeneous cosmologies. There has been extensive criticism of the BKL conjecture over the years (see, for example, [21.29]). One major concern has been the apparent dependence of their approach on the existence of an appropriate separation or slicing of the spacetime into space and time. Whether or not it is possible to make the necessary choice of time variable in all cases is not yet known. In support of the BKL conjecture, as we shall see below, are numerical simulations that appear to show the dominance of either Kasner-like or Mixmaster-like behavior at each spatial point [21.30]. Of course, such simulations also make a choice of spacetime slicing. Recent numerical studies [21.31] indicate that there may be phenomena in the approach to the singularity that do not align with the BKL conjecture. Later, we shall discuss frameworks that have been developed to allow a rigorous formulation of the BKL conjecture and to offer a possible path toward a proof. See [21.32, 33] and the references therein. A (toy) realization of the BKL conjecture is shown in Fig. 21.11 where a different Mixmaster simulation has been run at each spatial point. The artificial spatial dependence was induced by slowly varying the initial Kasner-circle momenta over the spatial grid.

21.3.2 Method of Consistent Potentials

The method of consistent potentials (MCP) was employed originally by *Moncrief* and *Grubišić* [21.34] to argue for the qualitative behavior of spatially inhomogeneous cosmologies. MCP may also be regarded as a simplification of BKL's arguments. To employ MCP, one first identifies a Kasner-like solution in the variables of the model at each spatial point. The remaining terms, usually from the spatial scalar curvature, may be treated as potentials. If, as the model evolves toward the singularity, the potential terms become exponentially small, the model becomes asymptotically Kasner-like. In spatially homogeneous models, one literally constructs the Kasner solution and compares the asymptotic behavior

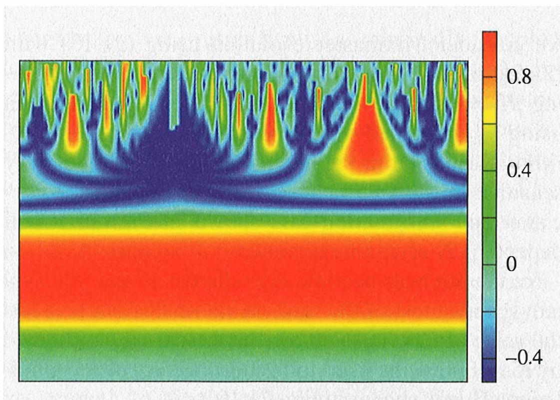


Fig. 21.11 A realization of the BKL conjecture. Mixmaster simulations as in [21.16] were begun with slowly varying initial conditions along the horizontal spatial grid to create fictitious spatial dependence in one direction. This caused the initial u-parameter to vary over the grid and to create bounces at different times at different *spatial* points. This in turn created ever smaller spatial structure. The evolution toward the singularity is upward. The variable plotted is $\beta_+/|\Omega|$ *with arbitrary color scale*

to it. In the spatially inhomogeneous case, the parameters of the Kasner-like solution depend on space but not on time. We shall refer to the Kasner-like asymptotic behavior as AVTD. The term *quiescent singularity* has also been used.

To illustrate MCP, consider the Taub model's Hamiltonian (21.25). The Kasner solution (21.23), when substituted in the exponential $e^{4\Omega-8\beta_+}$, yields the conditions for growth or decay of that term. We find

$$e^{4\Omega-8\beta_+} \to e^{4\Omega(1+2v_+)} . \quad (21.41)$$

This shows that if the velocity parameter $v_+ < -\frac{1}{2}$, the exponential will grow as $\Omega \to -\infty$, the direction of the singularity. Otherwise, the potential will be exponentially small. Now consider the bounce rule (21.29) relating v'_+ for the outgoing Kasner to the ingoing v_+. It is easy to see that, if $v_+ < -\frac{1}{2}$, $v'_+ > -\frac{1}{2}$, and no further bounces will occur. While the Kasner solution epitomizes VTD, the Taub model has at most one bounce making it AVTD. A realization of this behavior in the Taub model is shown in Fig. 21.12. Note that the potential is significant only when v_+ is in the range to cause a bounce.

If we extend the Taub MCP analysis to the Mixmaster model, we find that the three Mixmaster potentials

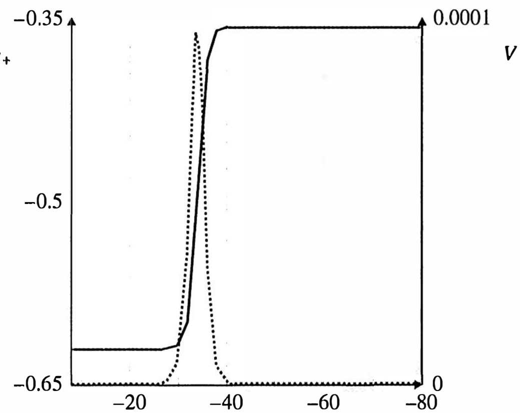


Fig. 21.12 Illustration of MCP. The approach to the singularity in the Taub model (as $\Omega \to -\infty$) shows an initial Kasner solution described by $v_+ < -\frac{1}{2}$ (*left vertical scale*). The trajectory then bounces off the growing potential $V = e^{4\Omega-8\beta_+}$ (*right vertical scale*) into a second Kasner regime with $v_+ > -\frac{1}{2}$ while the potential becomes exponentially small. The trajectory is shown in *brown* and the potential in *light grey*

cover the full range of $v_\pm$. There is no value of $v_\pm$ (except those corresponding to Taub initial data) where all the potential terms remain exponentially small. Thus, there is no last bounce. The Mixmaster model is not AVTD. We shall call collapsing spacetimes with no last bounce Mixmaster-like. Note that when a minimally coupled scalar field is added, collapsing Bianchi type IX models are AVTD.

MCP is easily applied to spatially inhomogeneous models (see subsequent sections) when Einstein's equations may be derived from a Hamiltonian. The Hamiltonian formulation of general relativity developed by Arnowitt, Deser, and Misner (ADM) (see [21.2] and elsewhere in this volume) rewrites Einstein's equations as derivable from the variation of the constraints with suitable lapse and shift. Schematically, the Hamiltonian constraint takes the form ([21.2] p. 465)

$$H = \frac{N}{\sqrt{^3g}}\left[\pi^{ij}\pi_{ij} - \frac{1}{2}\pi^2 - {}^3g({}^3R)\right] + N_i\nabla_j\left(\frac{\pi^{ji}}{\sqrt{^3g}}\right), \quad (21.42)$$

where the π_{ij} (with trace π) are related to the time derivatives of the spatial metric ${}^3g_{ij}$, ${}^3R$ is the spatial scalar curvature, the covariant derivative is formed from ${}^3g$, and N, N_i are the lapse and shift. In general, the kinetic part of the Hamiltonian constraint may be used to develop a Kasner-like solution at every spatial point. MCP is then implemented by checking the behavior of the spatial scalar curvature term in the approach to the singularity. Specific examples will be given later. It must be emphasized that MCP is a heuristic approach. It is possible to conclude from it that, for a given spatially inhomogeneous cosmology, the approach to the singularity is or is not locally Kasner-like (or AVTD). One cannot conclude that any non-Kasner-like behavior is, in fact, Mixmaster-like.

21.3.3 Mathematical, Heuristic, and Numerical Approaches for Specific Spacetimes

As mentioned previously, the mathematics of Mixmaster dynamics is highly nontrivial. Proofs exist only in the spatially homogeneous case. The apparently AVTD spacetimes based on MCP have been analyzed mathematically. Several methods have been employed. One was developed originally by Kichenasamy and applied to a variety of cosmological spacetimes by himself and

many others. For a review, see [21.35]. The basic idea is to choose variables to allow an asymptotic expansion starting at the singularity and proceeding away from it. One can then prove existence and other properties based on the expansion. For example, one can prove the existence of an open set of solutions to the Gowdy model Einstein equations with the AVTD property almost everywhere. In this section, we shall consider several examples of spatially inhomogeneous, (mostly) vacuum, cosmological spacetimes, how MCP may be applied, what the simulations show, and any mathematical statements that can be made.

The Gowdy Model on $T^3 \times R$

Gowdy discovered that if one interchanges the role of time and space in Einstein–Rosen waves, the resultant spacetimes may be interpreted as spatially inhomogeneous, vacuum cosmologies with spatial topologies T^3, S^3, and $S^2 \times S^1$ [21.36]. (In the first case, the topology may equally be taken to be R^3.) The approach to the singularity has been studied extensively in the first case (see below) and also for the last case [21.37]. The simplest spatially inhomogeneous cosmology is the Gowdy spacetime on $T^3 \times R$. The metric variables depend only on time τ and the periodic spatial coordinate θ. The model has two spatial Killing fields in the directions orthogonal to θ. In convenient variables, the metric is given by [21.18, 38]

$$\begin{aligned} \mathrm{d}s^2 &= \mathrm{e}^{\lambda/2}\mathrm{e}^{\tau/2}(-\mathrm{e}^{-2\tau}\,\mathrm{d}\tau^2 + \mathrm{d}\theta^2) \\ &\quad + \mathrm{e}^{-\tau}\left[\mathrm{e}^P\,\mathrm{d}\sigma^2 + 2\mathrm{e}^P Q\,\mathrm{d}\sigma\,\mathrm{d}\delta \right. \\ &\quad \left. + (\mathrm{e}^P Q^2 + \mathrm{e}^{-P})\,\mathrm{d}\delta^2\right] , \end{aligned} \tag{21.43}$$

where λ, P, and Q are periodic functions of θ and evolve in τ. The time choice is called *areal* and has been arranged so that the spatial volume $\to 0$ as $\tau \to \infty$. Einstein's equations for this model split into two groups. The first consists of nonlinearly coupled wave equations for P and Q (where $,_a = \partial/\partial a$)

$$P,_{\tau\tau} - \mathrm{e}^{2\tau}P,_{\theta\theta} = \mathrm{e}^{2P}\left(Q,_\tau^2 - \mathrm{e}^{2\tau}Q,_\theta^2\right), \tag{21.44}$$

$$Q,_{\tau\tau} - \mathrm{e}^{2\tau}Q,_{\theta\theta} = -2\left(P,_\tau Q,_\tau - \mathrm{e}^{2\tau}P,_\theta Q,_\theta\right). \tag{21.45}$$

The second contains the Hamiltonian and θ-momentum constraints, which can be expressed respectively as equations for $\lambda,_\tau$ and $\lambda,_\theta$ in terms of P and Q

$$\lambda,_\tau = -\left[P,_\tau^2 + \mathrm{e}^{-2\tau}P,_\theta^2 + \mathrm{e}^{2P}(Q,_\tau^2 + \mathrm{e}^{-2\tau}Q,_\theta^2)\right], \tag{21.46}$$

$$\lambda,_\theta + 2(P,_\theta P,_\tau + \mathrm{e}^{2P}Q,_\theta Q,_\tau) = 0 . \tag{21.47}$$

This means that λ may be constructed after a solution for P and Q has been obtained. This decoupling is an enormous simplification both numerically and mathematically. One may interpret these equations as follows: P and Q are related, respectively, to the amplitudes of the $+$ and $\times$ polarizations of the gravitational waves, while λ controls the background spacetime. The volume density $\sqrt{^3g}$ depends only on λ (and τ). The separability of the equations for P, Q from that for λ makes it convenient to define a Hamiltonian which yields the equations for P and Q

$$\begin{aligned} H &= \frac{1}{2}\int_0^{2\pi} \mathrm{d}\theta\,\left(\pi_P^2 + \mathrm{e}^{-2P}\pi_Q^2\right) \\ &\quad + \frac{1}{2}\int_0^{2\pi} \mathrm{d}\theta\,\left[\mathrm{e}^{-2\tau}\left(P,_\theta^2 + \mathrm{e}^{2P}Q,_\theta^2\right)\right] \\ &= H_K + H_V . \end{aligned} \tag{21.48}$$

Note that this Hamiltonian, H, is *not* the Hamiltonian constraint.

Polarized Gowdy T^3 models [21.39] are obtained by setting Q and π_Q equal to zero, an initial condition that is preserved by the evolution equations. The remaining wave equation for P has the explicit solution, written as a Fourier series

$$P = P_0 + \pi_P^0\tau + \sum_{n=1}^{\infty}\zeta_n Z_0(n\mathrm{e}^{-\tau})\cos(n\theta + \phi_n) , \tag{21.49}$$

where Z_0 is any zero-order Bessel function and ζ_n and ϕ_n are constants. The *zero mode* in the solution is just the Kasner solution (21.23) in these variables, while (21.43) with $Q = 0$ and without θ dependence is just a rewriting of the Kasner metric. As $\tau \to \infty$, (21.49) becomes (for a generic Bessel function of order zero)

$$P \approx P_0 + \left[\pi_P^0 + \sum_{n=1}^{\infty}\bar{\zeta}_n\cos(n\theta + \phi_n)\right]\tau . \tag{21.50}$$

The polarized Gowdy model has the asymptotic solution $P \approx v(\theta)\tau$ as $\tau \to \infty$, where $v(\theta)$ is the term in brackets in (21.50). In addition, λ may be constructed from (21.46) restricted to the polarized case. In the limit as $\tau \to \infty$, we obtain $\lambda \approx -v^2(\theta)\tau$, where $v(\theta)$ is the same as for P. Clearly this approximate solution

represents a different Kasner solution at every spatial point. This is clearly the VTD solution and demonstrates that the polarized Gowdy model is AVTD. The A appears because the local Kasner behavior is only valid as the singularity is approached. The mathematics of this model's approach to the singularity is well understood.

The generic Gowdy model also has a VTD solution. Variation of the Hamiltonian (21.48) keeping only the terms in H_K yields

$$
\begin{aligned}
P &= P_0 + \ln(\cosh v\tau + \cos\psi \sinh v\tau) \to v\tau \\
&\quad \text{as} \quad \tau \to \infty\,, \\
Q &= Q_0 + \frac{\mathrm{e}^{-P_0}\sin\psi \tanh v\tau}{1+\cos\psi\tanh v\tau} \to Q_\infty \\
&\quad \text{as} \quad \tau \to \infty\,, \\
\pi_P &= v\frac{\tanh v\tau + \cos\psi}{1+\cos\psi\tanh v\tau} \to v \quad \text{as} \quad \tau\to\infty\,, \\
\pi_Q &= \mathrm{e}^{P_0} v \sin\psi \equiv \pi_Q^0\,,
\end{aligned}
\tag{21.51}
$$

with a different solution at each spatial point (where the *constants* v, ψ, P_0, and Q_0 depend on θ but not on τ). We shall focus on the limit as $\tau \to \infty$ given here. Note that there is an overall factor of $\mathrm{e}^{-2\tau}$ in front of H_V in (21.48). This suggests, but does not prove, that the influence of the terms containing spatial derivatives will decrease as $\tau \to \infty$. It is important to note that the Hamiltonian H_K from (21.48) contains a potential-like term $V_1 = \pi_Q^2 \mathrm{e}^{-2P}$ that appears in the generic (unpolarized) case. It can be shown [21.40] that this term can be removed by a transformation of variables. However, the formulation used here yields a clearer MCP analysis.

Unlike in the polarized case, the potentials

$$
V_1 = \mathrm{e}^{-2P}\pi_Q^2\,; \quad V_2 = \mathrm{e}^{2P-2\tau} Q_{,\theta}^{\,2}
\tag{21.52}
$$

are present and affect the dynamics. Invocation of MCP by substituting the limiting VTD solution

$$
\begin{aligned}
P &= v(\theta)\tau\,; \quad \pi_P = \pi_P^0(\theta)\,; \\
Q &= Q_0(\theta)\,; \quad \pi_Q = \pi_Q^0(\theta)
\end{aligned}
\tag{21.53}
$$

into the potentials (21.52) yields the conditions that V_1 will become exponentially small if $v(\theta) > 0$, while V_2 will become exponentially small if $v(\theta) < 1$. Both conditions are satisfied if $0 < v(\theta) < 1$. However, the initial conditions (constructed to satisfy the momentum constraint) are not required to yield v within that range. As shown in Fig. 21.13 and as discussed in [21.38], just as in MSS, the potential terms act as walls causing the system point to bounce. Conservation of momentum yields the bounce laws $v \to -v$ off V_1 and $v \to 2-v$ off V_2. Successive bounces drive v into the range $(0, 1)$ and no further bounces occur. Thus the MCP prediction is that generic Gowdy models are AVTD in their approach to the singularity.

The mathematical results agree with this prediction. Kichenassamy and others [21.35] have shown that an open set of AVTD solutions exist in the vicinity of the singularity. Ringström has gone further to show that one can connect such behavior to the initial data away from the singularity [21.41].

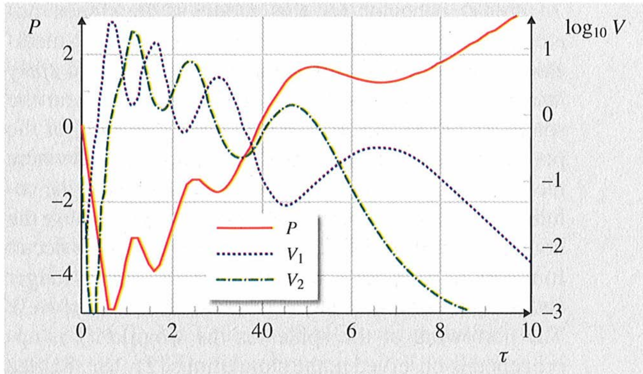


Fig. 21.13 Illustration of MCP at a typical spatial point in a Gowdy T^3 simulation. P (*red solid line*) initially has $|v(\theta)| > 1$. Subsequent bounces off V_1 (*dotted blue line*) and V_2 (*dash-dot green line*) drive $v(\theta)$ into the range $(0, 1)$. A version of this figure appeared in [21.38]

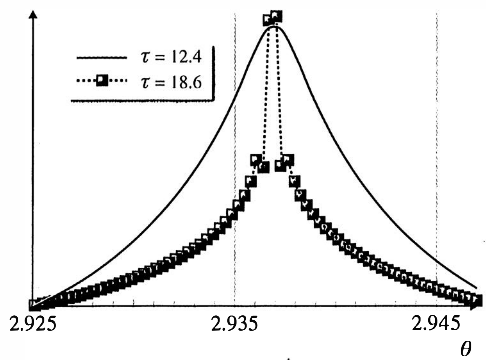


Fig. 21.14 The evolution of a spike in P shown at subsequent times (for τ increasing toward the singularity) in the simulation of a Gowdy T^3 model. The spike becomes ever narrower as bounces occur at θ values ever closer to the nongeneric point at the centroid of the spike. A version of this figure appeared in [21.38]

AVTD behavior for this model is thus seen numerically, understood via MCP, and confirmed mathematically. However, the simulations also reveal *spiky features*. These may be understood as arising near the spatial points where the coefficients π_Q^2 and Q_θ^2 of the potential terms (21.52) vanish. The actual points where they vanish are a set of measure zero in θ and in the solution. In the MCP picture, the spikes occur because the bounces to drive v into the allowed range do not occur. In the vicinity of these special points θ_n it takes longer and longer for the bounce to occur as $|\theta - \theta_n| \to 0$. The narrowing of the spikes as the singularity is approached is observed in the simulations [21.18, 38]. See Fig. 21.14 for an example.

Gowdy spikes are, in fact, completely understood [21.40]. The V_1-related spikes can be transformed away. Analytic expressions, called *spike solutions*, can be formulated for the V_2-related spikes. In addition, the spike solutions may be used to explain the evolution of $v(\theta)$ in the vicinity of the spike.

Generic T^2-Symmetric Models

However, the Gowdy models are not the most general T^2-symmetric vacuum spacetimes. (These spacetimes are also called G_2-symmetric after their two-dimensional spatial isometry group.) One can add additional off diagonal ($\theta - x$ and $\theta - y$) metric components to (21.43). Einstein's equations severely restrict the functional form obtained from the more general metric to the addition of two spacetime-independent *twist* constants. These may be reduced to a single twist constant κ without loss of generality. Details and some mathematical results related to the existence of particular choices of time coordinate may be found, e.g., in [21.42]. Details of this class of models, especially as presented here may be found in [21.43]. Without loss of generality, the metric may be written as

$$\begin{aligned}
ds^2 &= -e^{(\lambda-3\tau)/2} d\tau^2 + e^{(\lambda+\mu+\tau)/2} d\theta^2 \\
&\quad + e^{P-\tau} \left[d\sigma + Q d\delta \right. \\
&\quad \left. + \left(\int^\tau (Q\Theta) - Q \int^\tau \Theta \right) d\theta \right]^2 \\
&\quad + e^{-P-\tau} \left[d\delta - \left(\int^\tau \Theta \right) d\theta \right]^2 ,
\end{aligned} \tag{21.54}$$

where

$$\Theta = \kappa e^{\mu/4} e^{(\lambda+2P+3\tau)/2} . \tag{21.55}$$

Einstein's equations may be obtained by variation of the Hamiltonian density

$$\begin{aligned}
H &= \frac{1}{4\pi_\lambda} \left\{ \pi_P^2 + e^{-2P}\pi_Q^2 + e^{-2\tau} (\partial_\theta P)^2 \right. \\
&\quad \left. + e^{2(P-\tau)} (\partial_\theta Q)^2 \right\} \\
&\quad + \kappa^2 \pi_\lambda e^{(\lambda+2P+3\tau)/2} ,
\end{aligned} \tag{21.56}$$

with

$$\pi_\lambda - \frac{1}{2} e^{\frac{\mu}{4}} = 0 , \tag{21.57}$$

and subject to the momentum constraint

$$\pi_P P_{,\theta} + \pi_Q Q_{,\theta} + \pi_\lambda \lambda_{,\theta} = 0 . \tag{21.58}$$

The Hamiltonian constraint is the equation for $\lambda_{,\tau}$ obtained from the variation of (21.56) with respect to π_λ. The Gowdy model is recovered for $\pi_\lambda = \frac{1}{2}$ and $\kappa = 0$. Note that, in contrast to the Gowdy models, the wave amplitudes P and Q are no longer decoupled from a constituent equation for λ. We now have three potentials

$$\begin{aligned}
V_1 &= \frac{e^{-2P}\pi_Q^2}{4\pi_\lambda} ; \\
V_2 &= \frac{e^{2(P-\tau)}(Q_{,\theta})^2}{4\pi_\lambda} ; \\
V_3 &= \kappa^2 \pi_\lambda e^{(\lambda+2P+3\tau)/2} .
\end{aligned} \tag{21.59}$$

The VTD solution is, as $\tau \to \infty$, $\{Q, \pi_P, \pi_Q, \pi_\lambda\} \to \{Q_0(\theta), \pi_P^0(\theta), \pi_Q^0(\theta), \pi_\lambda^0(\theta)\}$ and, for

$$w = \frac{\pi_P^0}{2\pi_\lambda^0} \tag{21.60}$$

the limiting behavior as $\tau \to \infty$ is

$$\begin{aligned}
P &\to w\tau , \\
\lambda &\to -w^2\tau .
\end{aligned} \tag{21.61}$$

Application of MCP to these models yields the Gowdy conditions that V_1 is exponentially small if $w > 0$, that

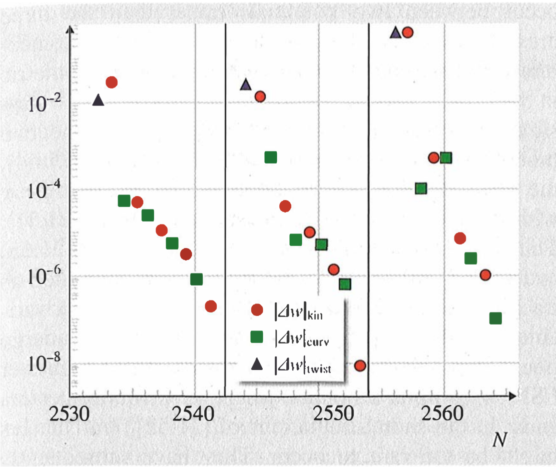


Fig. 21.15 Validation of the bounce laws for generic T^2-symmetric models. Simulation data at three adjacent spatial points are shown. The label N numbers subsequent bounces at a given spatial point and then switches to the next, adjacent point. The *vertical scale* shows the difference between the prediction from any of the bounce laws and from the simulation for the $N+1$-st value of w. The different bounce laws are color coded. *Alternating red dots* and *green squares* indicate that alternate bounces off V_1 and V_2 from (21.59) become an increasingly good description of the behavior. The initial data were arranged to begin each sequence with a twist bounce off V_3 (*plus in square*). A version of this figure appeared in [21.43]

V_2 is exponentially small if $w < 1$ and a new condition that V_3 is exponentially small as $\tau \to \infty$ if $w < -1$ or $w > 3$ but *not* if $0 < w < 1$. Thus, according to the MCP, there is no value of w where bounces would cease and the models are not AVTD. The bounce laws for each potential are given in Table III in [21.43]. The bounce laws are compared to typical behavior at given spatial points and seen to describe the behavior to high accuracy as shown in Fig. 21.15 [21.43]. Whether this corresponds to local Mixmaster dynamics is not entirely clear since the connection between the observed bounce laws and the BKL map for Mixmaster dynamics has not been made. See, however, [21.44] for a discussion of an indirect connection.

Unlike the Gowdy model spikes, spiky features in these generic T^2-symmetric models are not at fixed spatial points but rather move around and disappear and reappear during the evolution. *Lim* et al. [21.31] have shown that the simulations of these models contain embedded recurrent spikes described by an explicit function. They argue that this goes beyond BKL's conjecture. See also [21.45] for a more detailed treatment. However, BKL did not explicitly state their conjecture so that one might be able to relate the recurrent spikes to, e.g., special values of the parameter u (see (21.33)). While the T^2-symmetric models have been studied mathematically, these analyses do not address the nature of the approach to the singularity except in the case where a scalar field is added to suppress the oscillations [21.46].

The approach to the singularity of polarized T^2-symmetric models was studied recently by *Ames* et al. [21.47]. Application of MCP suggests that these models are AVTD, since bounces off V_1 will occur if $w < 0$ and off V_3 if $-1 < w < 3$. Thus, MCP predicts that bounces will not occur if $w > 3$. *Ames* et al. use Fuchsian methods to prove that AVTD solutions exist in these models.

$U(1)$ Symmetric Cosmologies

If the symmetry group is reduced from G_2 to $U(1)$ – that is, there is now only one spatial Killing field – we obtain a class of models described by the metric

$$\mathrm{d}s^2 = e^{-2\varphi}\left[-e^{2\Lambda}\,\mathrm{d}\tau^2 + e^{\Lambda} e_{ab}(x,z)\,\mathrm{d}\xi^a\,\mathrm{d}\xi^b\right] + e^{2\varphi}\left(\mathrm{d}\xi^3 + \beta_a\,\mathrm{d}x^a\,\mathrm{d}\tau\right)^2 , \tag{21.62}$$

where $a, b = 1, 2$ and φ, Λ, x, z, and β_a depend on spatial variables ξ_1, ξ_2, and time τ. The explicit form of e_{ab} is given in [21.48, 49] as is the discussion of a canonical transformation to replace the twists β_a with a single twist potential ω. The following details for this class of models may be found in [21.50]. Einstein's equations are obtained from variation of the Hamiltonian density

$$\begin{aligned}\mathcal{H} &= \left(\tfrac{1}{8}p_z^2 + \tfrac{1}{2}e^{4z}p_x^2 + \tfrac{1}{8}p^2 + \tfrac{1}{2}e^{4\varphi}r^2 - \tfrac{1}{2}p_\Lambda^2\right) \\ &\quad + (e^{\Lambda}e^{ab})_{,ab} - (e^{\Lambda}e^{ab})_{,a}\Lambda_{,b} \\ &\quad + e^{\Lambda}\left[(e^{-2z})_{,u}x_{,v} - (e^{-2z})_{,v}x_{,u}\right] \\ &\quad + 2e^{\Lambda}e^{ab}\varphi_{,a}\varphi_{,b} + \tfrac{1}{2}e^{\Lambda}e^{-4\varphi}e^{ab}\omega_{,a}\omega_{,b} ,\end{aligned} \tag{21.63}$$

where $\mathcal{H} = 0$ is the Hamiltonian constraint and p_φ, r, p_Λ, p_z, and p_x are canonically conjugate to φ, ω, Λ, z, and x. The VTD solution in these variables is

$$\begin{aligned}& z = -v_z\tau , \quad x = x_0 , \quad p_z = -4v_z , \\ & p_x = p_x^0 , \quad \varphi = -v_\varphi\tau , \\ & \omega = \omega_0 , \quad p = -4v_\varphi , \quad r = r^0 , \\ & \Lambda = \Lambda_0 + (2 - v_\Lambda)\tau , \quad p_\Lambda = v_\Lambda ,\end{aligned} \tag{21.64}$$

where ν_z, ν_φ, x_0, p_x^0, ω_0, r^0, Λ_0, and $\nu_\Lambda > 0$ are functions of ξ_1 and ξ_2 but independent of τ. (The sign of ν_Λ is fixed to ensure collapse.)

Polarized $U(1)$-symmetric models ($\omega = 0 = r$) have been examined both numerically [21.49] and analytically, where the case has been made that the singularity is AVTD. See, for example, [21.51] and the references therein for the mathematical results.

The generic (vacuum) $U(1)$ models are not AVTD according to MCP and numerical simulations. To date, the simulations have been too crude for detailed studies since too few Mixmaster-like bounces occur before the codes fail. MCP predicted that bounces would occur (in these variables) between two potentials [21.50]. We first notice that the Gowdy-like terms [21.18]

$$V_z = \tfrac{1}{2}p_x^2 e^{4z}\,, \quad V_1 = \tfrac{1}{2}r^2 e^{4\varphi}\,, \tag{21.65}$$

in $\mathcal{H}_K$ become, in the limit of $\tau \to \infty$, upon substitution of (21.64)

$$V_z \to \tfrac{1}{2}p_x^2 e^{-4\nu_z\tau}\,, \quad V_1 \to \tfrac{1}{2}r^2 e^{-4\nu_\varphi\tau}\,, \tag{21.66}$$

and are exponentially small only if $\nu_z > 0$ and $\nu_\varphi > 0$. (As in the Gowdy case [21.38], nongeneric behavior can arise at isolated spatial points where p_x and/or r vanish.) The remaining term is

$$V_2 = \frac{1}{2}e^{-2\tau+\Lambda}e^{-4\varphi}e^{ab}\omega_{,a}\,\omega_{,b}\,, \tag{21.67}$$

which becomes upon substitution of (21.64)

$$V_2 \approx F(x, \nabla\omega)e^{(-\nu_\Lambda+2\nu_z+4\nu_\varphi)\tau}\,, \tag{21.68}$$

where F includes all terms in the coefficient. The coefficients of τ in (21.66) and (21.68) are restricted by the VTD form of the Hamiltonian constraint (as $\tau \to \infty$)

$$\mathcal{H}^0 \approx -\tfrac{1}{2}\nu_\Lambda^2 + 2\nu_z^2 + 2\nu_\varphi^2 \approx 0\,, \tag{21.69}$$

obtained by substitution of (21.64) into (21.63). As discussed in [21.49], (21.69) implies that $\nu_\Lambda > 2\nu_z$, so that (21.66) decays exponentially for $\nu_z > 0$ for any ν_φ. On the other hand, for V_2 to become exponentially small with ν_z and $\nu_\varphi > 0$, we require $\nu_\Lambda^2 > (2\nu_z + 4\nu_\varphi)^2$, which is inconsistent with (21.69). Since there is no way to make V_1 and V_2 both exponentially small with the same value of ν_φ, the MCP predicts that either V_1 or V_2 will always grow exponentially. (Again, nongeneric behavior can result at isolated spatial points where the coefficient of V_2 happens to vanish.)

One might wonder if the $U(1)$ models actually show local Mixmaster dynamics given that bounces occur between two potentials rather than the three present in Bianchi IX models. An interesting conjecture can be found by writing the Bianchi IX metric in the $U(1)$ variables. For this purpose, it is possible to neglect the different topologies of the two model classes. As is discussed in detail in [21.44], the $U(1)$ variable φ may be identified at any time τ with the largest of the Mixmaster LSFs from (21.31). Thus the observed bounces off V_1 and V_2 from (21.65) and (21.67) in the $U(1)$ models may be identified as tracking all the Mixmaster bounces. The $U(1)$ variable z, on the other hand, was not observed to undergo bounces in [21.52]. Yet, in terms of the Mixmaster LSFs, z exhibits a bounce whenever a Mixmaster era ends. In the short simulations of [21.52], era changes might be too rare to occur. They have subsequently been sought by exploring a variety of initial conditions in the hope that a short era would occur at one or more spatial points. A possible example of a z-bounce is shown in Fig. 21.16. Longer simulations, if possible, could reveal more structure related to this phenomenology.

There are no mathematical results for generic $U(1)$ models relating to the nature of the approach to the singularity.

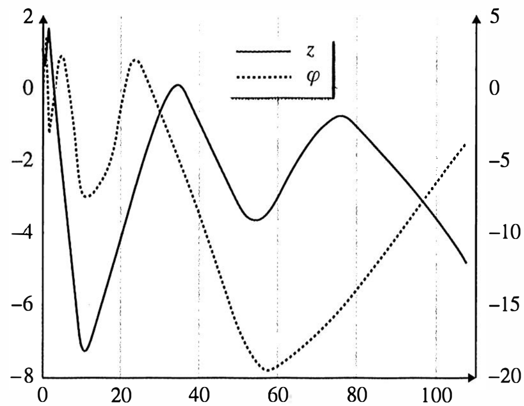


Fig. 21.16 A possible z-bounce in a $U(1)$-symmetric model. The $U(1)$ variables φ (*dotted light grey, right-hand scale*) and z (*brown, left-hand scale*) from (21.63) are shown versus time τ. The singularity is in the direction of increasing τ. Simulations in [21.52] do not show bounces in z. From the BKL map for u, it is easy to show that short eras with, say $3/2 < u < 2$ are likely to be followed by a second short era. Thus the two bounces seen in z are plausible. Note that the corresponding oscillations in φ are also shown

Generic Models with No Symmetries

The goal of the research described in this chapter is to elucidate the approach to the singularity in generic spacetimes. For generic spacetimes, there appears to be no particular advantage to using symmetry adapted variables as those used in the cases with symmetries described above. For simulations performed to date, it is convenient to use a modified version of the variables introduced by *Wainwright* and collaborators [21.25]. In terms of these variables, it is possible to construct invariants that reveal the nature of the dynamics. For example, in the variables of [21.33], one can identify the BKL parameter u in generic collapse of a vacuum spacetime with T^3 topology [21.53]. Figure 21.17 taken from [21.53] provides numerical evidence of local Mixmaster dynamics in generic collapse. In this simulation, u followed the BKL map (21.33) to some accuracy. However, the simulations had relatively low spatial resolution and further research in this direction is desirable.

On the other hand, if one considers a generic spacetime with $T^3 \times R$ topology with a scalar field, it can be proven that AVTD solutions exist on an open set close to the singularity [21.54]. The function of the scalar field is to suppress Mixmaster-like oscillations in the same way as in the spatially homogeneous case.

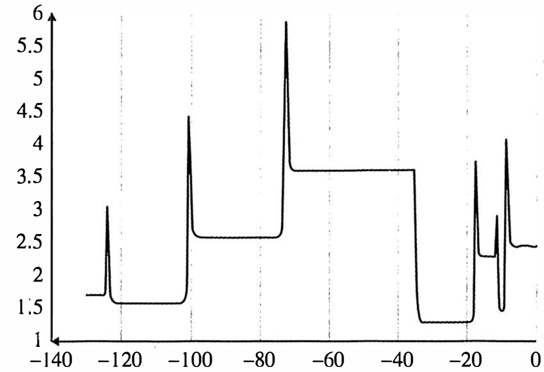


Fig. 21.17 The parameter u at a typical spatial point in a generic, collapsing, vacuum spacetime on $T^3 \times R$. This figure is taken from [21.53] where details may be found. Evolution toward the singularity is leftward. The apparent spikes occur during the bounces while the value of u is changing

21.4 Summary

In this chapter, we have focused on three classes of spatially homogeneous, vacuum cosmological spacetimes, described their behavior as dynamics in MSS, and characterized their approach to the singularity. Bianchi type I (Kasner) solutions formed the basic building block – the (fictitious) free particle in MSS. The next step, the Bianchi type II (Taub) solution, added a single potential wall in MSS. The dynamics could be completely understood through development of a bounce law relating the Kasner trajectory after the bounce to the one before. Since, in MSS, the Kasner models have no potential, they are called VTD. The Taub model has at most a single bounce and is thus asymptotically VTD or AVTD. Both these models approach the singularity to end in a curvature blow-up space-like singularity. The final complication arises in Bianchi type IX (Mixmaster) where one wall in MSS is replaced by three to form what appears to be an enclosed triangle. Numerical evidence and MCP indicate that these models continue to bounce forever in their approach to the singularity. Mathematical proofs now exist that demonstrate that the approach to the singularity in Bianchi type IX and other models with Mixmaster-like behavior, in fact, have this behavior all the way to the singularity ending in a curvature blow-up space-like singularity.

BKL long ago conjectured that the approach to the singularity in spatially inhomogeneous cosmologies followed the path of the spatially homogeneous models at every spatial point – the solutions were locally AVTD or Mixmaster-like at almost every spatial point. MCP and numerical studies have provided support for this picture in a variety of spacetimes. In cases where the conjectured and observed behaviors appear to be AVTD, mathematical support exists in the form of proofs that open sets of solutions have the required character in the vicinity of the curvature blow-up space-like singularity. The Mixmaster-like case is more complicated. The numerical evidence is supportive although non-Mixmaster-like behavior has been observed in the approach to the singularity of generic T^2-symmetric models. Mathematical proofs do not exist for non-AVTD spacetimes, although several groups are working on frameworks to allow a precise formulation of the BKL conjecture with the eventual possibility of proof – one way or the other.

21.5 Open Questions

We shall conclude this chapter with a discussion of open questions. The approach to the singularity in spatially homogeneous cosmologies appears to be under control both mathematically and numerically. There has been recent research in higher dimensions and in the influence of matter that behaves comparably to anisotropy energy (e.g., scalar fields with a variety of potentials [21.13, 20], cosmological constants, and, perhaps, dark energy or inflatons).

The spatially inhomogeneous models and the BKL conjecture are less well understood. In particular, there are no rigorous mathematical results for the nature of the approach to the singularity in models where one expects local Mixmaster dynamics (or something else which is not AVTD). Two competing frameworks have been developed. The first by *Damour* and collaborators [21.46] focuses on *billiards*, an approach to dynamical systems that can be studied as system points bouncing within sharp walls. They attempt to apply what is known mathematically about such systems to what essentially becomes a field of billiards. This is one way to formalize and make precise the BKL conjecture. In contrast, *Uggla* and collaborators [21.55] consider the extension of the Wainwright approach to spatially inhomogeneous models. This approach focuses on the Kasner circle with potential walls appearing (at every spatial point) as structures imposed on the Kasner circle at that spatial point. This approach also has the aim of development of a rigorous statement and analysis of the BKL conjecture. There has been some recent progress in rigorous statement of the BKL conjecture using these methods and in relating the two approaches [21.33]. So far, neither approach has achieved mathematical results for spatially inhomogeneous models that are not AVTD comparable to those found in the AVTD case. A recent development, motivated by loop quantum gravity, is a reformulation of the description of spatially inhomogeneous cosmological spacetimes to allow rigorous statement of the BKL conjecture [21.32].

It should be noted that, even for spatially inhomogeneous cosmologies with AVTD behavior, the theorems already proven are somewhat limited to existence of solutions with the desired behavior. Stronger results would be useful for more classes of spacetimes.

Of course, singularities also develop in noncosmological spacetimes, most notably in the interiors of black holes. It is still unclear what happens to the Cauchy horizon in the interiors of charged and rotating black holes and under what circumstances the asymptotic behavior resembles the approach to the space-like singularities discussed in this chapter. See [21.19] and elsewhere in this volume for further discussion of this topic.

A related issue is the phase-transition-like critical behavior that was first discovered numerically by Choptuik in the collapse of a spherically symmetric scalar field to form a black hole or to disperse. Similar behavior has been found in a variety of other systems. For reviews, see [21.19, 56]. It is also of interest to consider what happens in more than three spatial dimensions, particularly in models motivated by string theory and/or supersymmetry.

Finally, we note that collapsing gravitational systems that become singular in a finite time signal a failure of general relativity. It has long been hoped that these singularities can be removed when quantum mechanical effects (either through quantum field theory in curved spacetime or quantum gravity) are included. Progress on this topic is discussed elsewhere in this volume.

References

21.1 S.M. Carroll: *Spacetime and Geometry* (Addison Wesley, San Francisco 2004)

21.2 R.M. Wald: *General Relativity* (Univ. Chicago Press, Chicago 1984)

21.3 G.F.R. Ellis, M.A.H. MacCallum: A class of homogeneous cosmological models, Comm. Math. Phys. **12**, 108 (1969)

21.4 M.A.H. MacCallum: Anisotropic and inhomogeneous relativistic cosmologies. In: *General Relativity: An Einstein Centenary Survey*, ed. by S. Hawking, W. Israel (Cambridge Univ. Press, Cambridge, 1979) pp. 533–580

21.5 M.P. Ryan Jr., L.C. Shepley: *Homogeneous Relativistic Cosmologies* (Princeton Univ. Press, Princeton, 1975)

21.6 C.W. Misner: Minisuperspace, magic without magic. In: *J. A. Wheeler 60th Anniversary Volume*, ed. by J. Klauder (W.H. Freeman, San Francisco 1972) pp. 441–473

21.7 E. Kasner: Solutions of the Einstein equations involving functions of only one variable, Trans. Am. Math. Soc. **27**, 155–162 (1925)

21.8 V.A. Belinskii, E.M. Lifshitz, I.M. Khalatnikov: Oscillatory approach to the singularity point in rel-

ativistic cosmology, Sov. Phys. Usp. **13**, 745–765 (1971)
21.9 B.K. Berger: Comments on the computation of Liapunov exponents for the mixmaster universe, Gen. Relativ. Gravit. **23**, 1385–1402 (1991)
21.10 D.M. Eardley, E. Liang, R. Sachs: Velocity-dominated singularities in irrotational dust cosmologies, J. Math. Phys. **13**, 99–107 (1972)
21.11 A. Taub: Empty space-times admitting a three-parameter group of motions, Ann. Math. **53**, 472 (1951)
21.12 C.W. Misner: Mixmaster universe, Phys. Rev. Lett. **22**, 1071–1074 (1969)
21.13 R.T. Jantzen: Spatially homogeneous dynamics: A unified picture, Proc. Int. School Phys. 'Enrico Fermi', Course 86, Varenna, Italy, 1982 (North-Holland Elsevier, Amsterdam 1986) pp. 61–147
21.14 C.W. Misner: The mixmaster cosmological metrics, NATO ASI Ser. **332**, 317–328 (1994)
21.15 B.K. Berger: How to determine approximate mixmaster parameters from numerical evolution of Einstein's equations, Phys. Rev. D **49**, 1120–1123 (1994)
21.16 B.K. Berger, D. Garfinkle, E. Strasser: New algorithm for mixmaster dynamics, Class. Quantum Gravity **14**, L29–L36 (1997)
21.17 D.W. Hobill, A. Burd, A.A. Coley: Deterministic Chaos in General Relativity, NATO ASI Ser. **332** (1994)
21.18 B.K. Berger, V. Moncrief: Numerical investigations of cosmological singularities, Phys. Rev. D **48**, 4676–4687 (1993)
21.19 B.K. Berger: Numerical approaches to spacetime singularities, Living Rev, Relativity **5**, 1 (2002)
21.20 V.A. Belinskii, I.M. Khalatnikov: Effect of scalar and vector fields on the nature of the cosmological singularity, Sov. Phys. JETP **36**, 591–597 (1973)
21.21 B.K. Berger: Influence of scalar fields on the approach to the singularity in spatially inhomogeneous cosmologies, Phys. Rev. D **61**, 023508 (2000)
21.22 B.K. Berger: Why solve the Hamiltonian constraint in numerical relativity?, Gen. Relativ. Gravit. **38**, 625–632 (2006)
21.23 H. Ringström: Curvature blow up in Bianchi VIII and IX vacuum spacetimes, Class. Quantum Gravity **17**, 713–731 (2000)
21.24 M. Weaver: Dynamics of magnetic Bianchi VI_0 cosmologies, Class. Quantum Gravity **17**, 421–434 (2000)
21.25 J. Wainwright: A dynamical systems approach to Bianchi cosmologies: Orthogonal models of class A, Class. Quantum Gravity **6**, 1409 (1989)
21.26 H. Ringström: The Bianchi IX attractor, Ann. Henri Poincaré **2**, 405–500 (2001)
21.27 J.M. Heinzle, C. Uggla: A new proof of the Bianchi type IX attractor theorem, Class. Quantum Gravity **26**, 075015 (2009)
21.28 V.A. Belinskii, E.M. Lifshitz, I.M. Khalatnikov: A general solution of the Einstein equations with a time singularity, Adv. Phys. **13**, 639–667 (1982)
21.29 J.D. Barrow, F.J. Tipler: Analysis of the generic singularity studies by Belinskii, Khalatnikov, and Lifshitz, Phys. Rep. **56**, 371–402 (1979)
21.30 B.K. Berger, D. Garfinkle, J.A. Isenberg, V. Moncrief, M. Weaver: The singularity in generic gravitational collapse is spacelike, local, and oscillatory, Mod. Phys. Lett. A **13**, 1565–1574 (1998)
21.31 W.C. Lim, L. Andersson, D. Garfinkle, F. Pretorius: Spikes in the mixmaster regime of G_2 cosmologies, Phys. Rev. D **79**, 123526 (2009)
21.32 A. Ashtekar, A. Henderson, D. Sloan: A Hamiltonian formulation of the BKL conjecture, Phys. Rev. D **83**, 084024 (2011)
21.33 J.M. Heinzle, C. Uggla, N. Rohr: The cosmological billiard attractor, Adv. Theor. Math. Phys. **13**, 293–407 (2009)
21.34 B. Grubišić, V. Moncrief: Asymptotic behavior of the $T^3 \times R$ Gowdy space-times, Phys. Rev. D **47**, 2371–2382 (1993)
21.35 A.D. Rendall: Fuchsian methods and spacetime singularities, Class. Quantum Gravity **21**, S295–S304 (2004)
21.36 R.H. Gowdy: Gravitational waves in closed universes, Phys. Rev. Lett. **27**, 826 (1971)
21.37 D. Garfinkle: Numerical simulations of Gowdy spacetimes on $S^2 \times S^1 \times R$, Phys. Rev. D **60**, 104010 (1999)
21.38 B.K. Berger, D. Garfinkle: Phenomenology of the Gowdy model on $T^3 \times R$, Phys. Rev. D **57**, 4767–4777 (1998)
21.39 B.K. Berger: Quantum graviton creation in a model universe, Ann. Phys. **83**, 458–490 (1974)
21.40 A.D. Rendall, M. Weaver: Manufacture of Gowdy spacetimes with spikes, Class. Quantum Gravity **18**, 2959–2975 (2001)
21.41 H. Ringström: Asymptotic expansions close to the singularity in Gowdy spacetimes, Class. Quantum Gravity **21**, S305–S322 (2004)
21.42 B.K. Berger, P.T. Chruściel, J.A. Isenberg, V. Moncrief: Global foliations of vacuum spacetimes with T^2 isometry, Ann. Phys. **260**, 117–148 (1997)
21.43 B.K. Berger, J.A. Isenberg, M. Weaver: Oscillatory approach to the singularity in vacuum spacetimes with T^2 isometry, Phys. Rev. D **64**, 084006 (2001)
21.44 B.K. Berger: Hunting local mixmaster dynamics in spatially inhomogeneous cosmologies, Class. Quantum Gravity **21**, S81–S96 (2004)
21.45 J.M. Heinzle, C. Uggla, W.C. Lim: Spike Oscillations, Phys. Rev. D **86**, 104049 (2012)
21.46 T. Damour, M. Henneaux, A.D. Rendall, M. Weaver: Kasner-like behaviour for subcritical Einstein-matter systems, Ann. Henri Poincare **3**, 1049–1111 (2002)
21.47 E. Ames, F. Beyer, J. Isenberg, P.G. LeFloch: Quasilinear hyperbolic fuchsian systems and AVTD behavior in T^2-symmetric vacuum spacetimes, Annal. Henri Poincaré **14**, 1445–1523 (2012)

21.48 V. Moncrief: Reduction of Einstein's equations for vacuum space-times with spacelike $U(\mathbf{1})$ isometry groups, Ann. Phys. **167**, 118–142 (1986)

21.49 8.K. Berger, V. Moncrief: Numerical evidence that the singularity in polarized $U(1)$ symmetric cosmologies on $T^3 \times R$ is velocity dominated, Phys. Rev. D **57**, 7235–7240 (1998)

21.50 B.K. Berger, V. Moncrief: Evidence for an oscillatory singularity in generic $U(1)$ symmetric cosmologies on $T^3 \times R$, Phys. Rev. D **58**, 1–8 (1998)

21.51 Y. Choquet-Bruhat, J. Isenberg, V. Moncrief: Topologically general $U(1)$ symmetric Einstein spacetimes with AVTD behavior, Nuovo Cim. B **119**, 625–638 (2004)

21.52 B.K. Berger, V. Moncrief: Evidence for an oscillatory singularity in generic $U(1)$ symmetric cosmologies on $T^3 \times R$, Phys. Rev. D **58**, 064023 (1998)

21.53 D. Garfinkle: Numerical simulations of generic collapse, Phys. Rev. Lett. **93**, 161101 (2004)

21.54 L. Andersson, A.D. Rendall: Quiescent cosmological singularities, Commun. Math. Phys. **218**, 479–511 (2001)

21.55 C. Uggla, H. van Elst, J. Wainwright, G.F.R. Ellis: The past attractor in inhomogeneous cosmology, Phys. Rev. D **68**, 103502 (2003)

21.56 C. Gundlach, J.M. Martin-Garcia: Critical phenomena in gravitational collapse, Living Rev, Relativity **10**, 5 (2007)

Part D

Confront

Part D Confronting Relativity Theories with Observations

22. The Experimental Status of Special and General Relativity

Orfeu Bertolami, Jorge Páramos

In this contribution we assess the current experimental status of special and general relativity. Particular emphasis is put on putative extensions of these theories and on how these could be detected experimentally.

22.1 Introductory Remarks

Special relativity (SR) was proposed more than 100 years ago and has allowed for a profound change in our perspective of the fundamental building blocks of physics, namely space and time, which until then had been regarded as immutable and absolute.

The generalization of SR to encompass general coordinate transformations and, through the equivalence principle, to also incorporate gravity, has led to an inevitable connection to the mathematics of curved spaces, putting general relativity (GR) in a unique standing among physical theories; GR is a theory of space and time, thus setting the tools to describe the dynamics and the evolution of the Universe as a whole.

From the conceptual point of view, Relativity was a major step forward; the pressure to unravel putative extensions to this theory of gravity leads one to carefully test the foundational principles of SR and GR (see [22.1–3] and references therein).

22.2 Experimental Tests of Special Relativity

More than a century ago, Einstein put forward his revolutionary special theory of relativity (SR), so called because it accounted only for phenomena seen from inertial reference frames, which move with constant relative velocity. Although several reformulations have arisen in the intervening years, with added clarity and mathematical precision [22.4], Einstein resorted to two fundamental postulates in order to derive SR:

- The principle of relativity, which states that physical laws are independent of the inertial reference frame used to infer them.
- The constancy of the speed of light, which is always propagated in empty space at $c \approx 3\times10^8$ m/s, independently of the state of motion of its source.

Both postulates may be shown to lead to the concept of Lorentz invariance, i. e., that the laws of physics are

invariant with respect to the Lorentz transformations; if one takes two inertial frames S and S' with relative speed v in the x-axis, these amount to the well-known relations between time and space coordinates

$$x' = \frac{x - vt}{\sqrt{1 - \frac{v^2}{c^2}}}\,,$$
$$y' = y\,,$$
$$z' = z\,,$$
$$t' = \frac{t - \frac{vx}{c^2}}{\sqrt{1 - \frac{v^2}{c^2}}}\,. \tag{22.1}$$

These transformations were known to leave Maxwell's equations invariant, while the Galileo transformations, which leave mechanics invariant, did not. However, it was Einstein who first understood that they could not be framed in a classical, Newtonian worldview, accompanied by a suitable aether medium, but instead required a fundamental rethinking of the concepts of space and time. The eponymous experiments carried out by Michelson and Morley in 1887 where taken not only as a disproof of this *luminous aether*, but as an observational evidence for the constancy of the speed of light.

Notice that the above postulates do not make any claim concerning the equivalence between mass and energy, since they are of a kinematic (or geometric) nature alone. However, Einstein's derivation of this relationship resorts to a putative Lorentz invariance, used to establish the Lorentz invariant $(cp)^2 = E^2 - m^2c^4$ involving momentum, energy and rest mass.

The above introduction serves not only as a historical introduction, but helps to assess what are the most likely signals of Lorentz violation: privileged frame effects, variations in the speed of light [22.5], or failure of the Lorentz transformations altogether; other consequences include deviations from the $p^2 = E^2 - m^2c^4$ dispersion law, or maximum attainable speeds $c_i \neq c$ for different matter species.

In the realm of theories of gravity, competing theories to the *de facto* gold standard GR are usually experimentally assessed via the so-called PPN formalism [22.1], discussed later in this text: for now, it suffices to state that this formalism relies on an expansion of the dynamical metric field $g_{\mu\nu}$ in terms of suitable potentials, and the ensuing identification of a set of PPN parameters signaling deviation from GR. Since SR is restricted to inertial frames and negligible gravitational fields, thus assuming the a priori Minkowski metric $g_{\mu\nu} = \eta_{\mu\nu}$, such a tool is not valid when addressing the issue of Lorentz symmetry breaking in SR.

Nevertheless, one may resort to a similar expansion of some fundamental relation or quantity, with the expansion coefficients being related to alternative theories to SR, that breaks Lorentz invariance. The brief discussion above serves to better settle the three candidates that arise prominently: the Lorentz transformations themselves, the speed of light c, and the dispersion relation $p^2 = p^2(E, m, \chi)$ (where χ symbolizes additional properties or fields not present in the equivalent SR relation).

Each of these *test subjects* leads to a widely different formalism, which also reflects whether its motivation is one of the following:

- Kinematic, i. e., a relatively straightforward description of deviations from SR in the motion of massive bodies, propagation of light, causality, observability, light cone considerations, etc.
- Dynamical, in which case it attempts to formulate the intrinsic behavior of fields and fundamental equations in terms of Lorentz breaking quantities, thus allowing one to capture other relativistic behavior such as the clock rates of physical clocks (e.g., atomic clocks), light polarization effects, etc.

Given the variety of kinematic and dynamical formalisms available, it is somewhat difficult to compare them directly, either in terms of constraints to their defining observables, or when attempting to address a particular theoretical Lorentz breaking construction (see [22.6] for a discussion).

22.2.1 The Robertson–Sexl–Mansouri Formalism

Historically, this was the first attempt to put forward a formalism embodying the possibility of Lorentz symmetry breaking through the deviation of some free parameters from their SR values [22.7]; this is addressed by assuming that a privileged frame $\Sigma(T, \boldsymbol{X})$ exists (usually considered the cosmological frame, defined as that where the cosmic microwave background radiation appears isotropic and homogeneous at large scales), so that transforming from this to another inertial frame $S(t, \boldsymbol{x})$ with relative velocity $\boldsymbol{v}$ is achieved via

the deformed transformations

$$T=\frac{t-\boldsymbol{\epsilon}\cdot\boldsymbol{x}}{a}\,,\quad \boldsymbol{X}=\frac{\boldsymbol{x}}{d}-\left(\frac{1}{d}-\frac{1}{b}\right)\frac{\boldsymbol{v}\cdot\boldsymbol{x}}{v^2}\boldsymbol{v}+\frac{\boldsymbol{v}t}{a}\,. \tag{22.2}$$

Comparing this with (22.1), one finds that in SR

$$a=b^{-1}=\sqrt{1-\frac{v^2}{c^2}}\,,\quad d=1\,. \tag{22.3}$$

The vector $\boldsymbol{\epsilon}$, although not uniquely determined, does not add any further information concerning a putative breaking of Lorentz invariance, but reflects the chosen clock synchronization procedure: momentary external synchronization leads to $\boldsymbol{\epsilon}=\mathbf{0}$, while Einstein synchronization implies $\boldsymbol{\epsilon}=-a\boldsymbol{v}/b(1-v^2)$; slow transport of clocks leads to $(\nabla_v a)/b$, although additional complications due to clock rate variations between spacetime points can arise from dynamical effects on the clock mechanism (i. e., shifts in atomic transition frequencies). As a result, physical observables do not depend on the particular choice of $\boldsymbol{\epsilon}$, with the natural exception of those experiments which directly depend upon a particular synchronization method (see [22.1] for a discussion).

It is advantageous to resort to a set of numerical coefficients to parameterize the extent of violation of Lorentz symmetry, instead of using the functional form of a and b: since most conceivable experiments involve massive bodies endowed with nonrelativistic speeds, one may attain this by expanding these quantities around $(v/c)^2$ to second order, thus obtaining

$$\begin{aligned}
a(v)&=1+\left(\alpha-\tfrac{1}{2}\right)\left(\frac{v}{c}\right)^2+\left(\alpha_2-\tfrac{1}{8}\right)\left(\frac{v}{c}\right)^4,\\
b(v)&=1+\left(\beta+\tfrac{1}{2}\right)\left(\frac{v}{c}\right)^2+\left(\beta_2+\tfrac{3}{8}\right)\left(\frac{v}{c}\right)^4,\\
d(v)&=1+\delta\left(\frac{v}{c}\right)^2+\delta_2\left(\frac{v}{c}\right)^4,\\
\boldsymbol{\epsilon}&=(\epsilon-1)\left[1+\epsilon_2\left(\frac{v}{c}\right)^2\right]\boldsymbol{v}\,.
\end{aligned} \tag{22.4}$$

The choice of expansion coefficients is made so that, upon comparison with (22.3), one finds that SR yields all vanishing parameters except ϵ and ϵ_2, as discussed above; Einstein synchronization, the usual procedure followed in SR, also yields $\epsilon=\epsilon_2=0$.

Using the above expressions, one may derive a convoluted expression for the speed of light,

$$\begin{aligned}
c&=1-\epsilon\cos\theta\frac{v}{c}\\
&\quad-\left[\delta-\alpha+(\beta-\gamma+\epsilon^2)\cos^2\theta\right]\left(\frac{v}{c}\right)^2\\
&\quad+\{\beta-\alpha+\epsilon_2-\epsilon[2(\alpha+\delta)+\epsilon_2]\\
&\qquad-\epsilon[2(\beta-\delta)+\epsilon^2]\cos^2\theta\}\cos\theta\left(\frac{v}{c}\right)^3\\
&\quad+\Bigg(\delta_2-\alpha_2-\alpha\left(\tfrac{1}{2}+\delta-\alpha\right)\\
&\qquad+\Bigg\{\beta_2-\delta_2-\beta\left(\frac{1+3\beta}{2}+\alpha-3\delta+2\epsilon\right)\\
&\qquad\quad+\alpha[\delta+(2-3\epsilon)\epsilon]-3\delta\left(\frac{\delta}{2}-\epsilon^2\right)\\
&\qquad\quad+2(\epsilon-1)\epsilon\epsilon_2\\
&\qquad\quad+\left[3(\beta-\delta)\left(\frac{\beta-\delta}{2}+\epsilon^2\right)+\epsilon^4\right]\\
&\qquad\quad\times\cos^2\theta\Bigg\}\cos^2\theta\Bigg)\left(\frac{v}{c}\right)^4\,,
\end{aligned} \tag{22.5}$$

where θ is the angle between the velocity $\boldsymbol{v}$ of the frame of reference and the path of light; the independence of experiments not relying on a specific synchronization method on the related parameters ϵ and ϵ_2 becomes apparent if one computes the relative shift in the two-way speed of light $c_2(\theta,v)$,

$$\begin{aligned}
&\frac{c_2(\theta,v)}{c_2(0,v)}-1\\
&=\sin^2\theta\Bigg\{(\delta-\beta)\left(\frac{v}{c}\right)^2\\
&\quad\times\left[\frac{3\delta^2-\beta^2}{4}+\beta_2-\frac{\beta}{2}(1+\delta)\right.\\
&\quad\left.-\delta_2+\tfrac{3}{4}(\beta-\delta)^2\cos 2\theta\right]\left(\frac{v}{c}\right)^4\Bigg\}\,.
\end{aligned} \tag{22.6}$$

Similarly, the phase shift (not shown here for brevity), which can be measured by interferometry, is also independent on ϵ and ϵ_2 (see [22.6] for details).

An experimental determination of any nonvanishing parameters would immediately indicate that the underlying physical theory is not Lorentz invariant. However, second-order tests (i. e., obtained by disregarding terms $O(v^4)$ above) have yielded impressive bounds on these

quantities: the most recent Michelson–Morley experiment probing the dependence of the speed of light on its orientation with respect to a preferred frame yielded $(\beta-\delta) = (4\pm 8)\times 10^{-12}$ [22.8], while a modification of its setup (the Kennedy–Thorndike experiment) showed no signal of any effect of c on the velocity of the apparatus, $(\alpha-\beta) = -4.8(3.7)\times 10^{-8}$ [22.9]; finally, the most precise relativistic Doppler effect measurement has shown that time dilation as predicted by SR is valid down to a precision of $|\alpha| \leq 8.4\times 10^{-8}$ [22.10]. Thus, no violation of SR or any of its foundational principles has been detected so far.

22.2.2 The c^2 Formalism

Another formalism to address the possibility of breaking Lorentz invariance arises if one disregards the postulate of the constancy of the speed of light. (For clarity, one does not assume in this paragraph the natural system of units, in which $c = 1$.) This is best attained by resorting to the so-called $TH\epsilon\mu$ framework [22.11], an alternative to the PPN formalism when parameterizing gravity theories that deviate from GR [22.1].

This formalism is well suited to describe the interaction between charged particles in an external static and spherically symmetric gravitational field resulting from some metric theory of gravity: the field $T = g_{00}$ describes the temporal component of the metric $g_{\mu\nu}$, while isotropy allows one to express the spatial part as $g_{ij} = H\eta_{ij}$. The μ and ϵ parameters act as a generalization of the magnetic permeability μ_0 and electric permittivity ϵ_0 of a medium; depending on the underlying physical theory, these may depend on internal structure or the effect of other bodies.

The $TH\epsilon\mu$ formalism is able to signal deviations from metricity via a set of appropriately defined parameters

$$\begin{aligned} \Gamma_0 &= -c_0^2 \frac{\partial}{\partial U} \ln\left[\epsilon\sqrt{\frac{T}{H}}\right], \\ \Lambda_0 &= -c_0^2 \frac{\partial}{\partial U} \ln\left[\mu\sqrt{\frac{T}{H}}\right], \\ \Upsilon_0 &= 1 - \frac{T}{H}\epsilon\mu, \end{aligned} \quad (22.7)$$

which vanish if the EP is valid. More rigorously, the Einstein equivalence principle (EP), which comprises weak EP, local Lorentz invariance and local position invariance. In the above, $c_0 = \sqrt{T/H}$ is shown to be the limiting speed of material test particles; the latter contrasts with the speed of light $c = 1/\sqrt{\epsilon\mu}$, which follows the usual definition stemming from Maxwell's equations; however, both speeds c_0 and c allow for a spacetime and/or constitution dependency, that is, a nonconstant c breaks Einstein's second postulate of SR, while EP is broken if $c \neq c_0$ (even if $c_0 = const.$ is the same for all matter species).

Since SR is obtained by taking the flat spacetime limit of GR, one may extract a suitable formalism to address Lorentz symmetry breaking in negligible gravitational fields and inertial frames by considering the same limiting case of the $TH\epsilon\mu$ formalism: this is achieved by considering the c^2 formalism [22.1], and is attained by removing the spacetime dependence of the eponymous set of parameters, as if the dynamics of the gravitational field are disregarded. As a result, one is left with the possibility of deviations between c and c_0.

22.2.3 Modified Dispersion Relation

A more phenomenological, straightforward way of breaking Lorentz invariance is to assume that the SR dispersion relation $E^2 = p^2c^2 + m^2c^4$ is generalized to $E^2 = F(p,E)$, due to some underlying physical theory. Knowing the latter, one should also be able to establish the conservation laws for energy and momentum; in the absence of full knowledge of its inner workings, one may assume that both quantities are conserved, or resort to another phenomenological dependency for $\Delta E(p,E)$ and $\Delta p(p,E)$.

Since SR has withstood all tests so far, one knows that its dispersion relation must be a very good approximation, at least for the experimental regime available $v \ll c$. Thus, it is natural that the putative full dependence $E^2 = F(p,E)$ allows a Taylor expansion around $v = 0$, of the form

$$\begin{aligned} E^2 = m^2 + p^2 + M_P f_i^{(1)} p^i + f_{ij}^{(2)} p^i p^j + \frac{f_{ijk}^{(3)}}{M_P} p^i p^j p^k \\ + \dots, \end{aligned} \quad (22.8)$$

setting $c = 1$, for simplicity; the coefficients $f^{(n)}$ are dimensionless, having factored out the Planck mass M_P, the assumed scale at which relevant Lorentz symmetry breaking effects should arise due to some fundamental theory. These coefficients must be related to the underlying physical theory, and could be spacetime or position-dependent. More evolved modifications of the dispersion relation may arise if one assumes that spacetime is discretized [22.12, 13] or stochastic [22.14].

22.2.4 Dynamical Framework

The previous formalisms address the phenomenological implications of breaking Lorentz invariance via deformed relations for the coordinate transformations, the dispersion relation or the speed of light or limiting velocity of massive bodies. In stark contrast, one may conceive dynamical schemes that attempt to model Lorentz breaking extensions via an effective theory, valid at the low-energy, low-velocity regime.

Since one is dealing with the issue of testing Lorentz invariance at low energies, i. e., probing the validity of SR, gravity may be discarded from such an extension. The first setup in flat space is the minimal standard model extension (mSME) [22.15]; in this, the interactions of the standard model are enriched by a set of renormalizable Lorentz breaking operators involving fermions and the gauge bosons compatible with the internal gauge symmetry of QED. One may readily extend this to Yang–Mills theories, including models of the electromagnetic, weak and strong forces with an appropriate covering group. Gravity can also be included, as well as an embedding of our worldsheet into higher-dimensional braneworlds. Naturally, this dynamical framework encompasses the previously considered Robertson–Sexl–Mansouri formalism [22.16].

One focuses attention on the *minimal QED extension*, as it provides a sufficiently broad framework for the bulk of experimental tests of SR that involve electrons and light propagation. Further imposing $SU(2)$ gauge symmetry breaking, one can write the relevant additional terms to the Lagrangian density describing fermions and the electromagnetic field as [22.17]

$$\Delta \mathcal{L} = \frac{1}{2}\bar{\psi}\Gamma_\nu \overleftrightarrow{\partial}^\nu \psi - \bar{\psi}M\psi + \frac{1}{2}(\kappa_F)_{\alpha\beta\mu\nu}F^{\alpha\beta}F^{\mu\nu} , \tag{22.9}$$

where $F_{\mu\nu} = \partial_\mu A_\mu - \partial_\nu A_\mu$ is the usual field strength tensor. In the fermionic sector, one introduces a generalized mass term

$$M \equiv m + a_\mu\gamma^\mu + b_\mu\gamma_5\gamma^\mu + \frac{1}{2}H_{\mu\nu}\sigma^{\mu\nu} , \tag{22.10}$$

where m is the *bare* mass, as well as generalized gamma matrices

$$\begin{aligned}\Gamma_\nu \equiv\ & \gamma_\nu + c_{\mu\nu}\gamma^\mu + d_{\mu\nu}\gamma_5\gamma^\mu + e_\nu + if_\nu\gamma_5 \\ & + \frac{1}{2}g_{\alpha\mu\nu}\sigma^{\alpha\mu} ,\end{aligned} \tag{22.11}$$

where the a_μ, b_μ, $c_{\mu\nu}$, $d_{\mu\nu}$, e_ν, f_ν, $\sigma_{\mu\nu}$, and $H_{\mu\nu}$ are parameters that should arise from the underlying high energy theory. It is worthwhile to notice that if the breaking of Lorentz invariance is spontaneous, i. e., this is an exact symmetry of the latter, then these parameters are related to vacuum expectation values of Lorentz tensors and must be CPT-invariant. A toy model where a suitable number of vectors couple to the Ricci curvature introduces, when the former acquire a vacuum expectation value, spontaneous Lorentz symmetry breaking into the gravity sector and yields interesting astrophysical implications [22.18, 19]. Hermiticity of $\mathcal{L}$ also implies that they are real.

Dropping higher-order operators, which should not be as relevant in the low-energy limit, one expects a fermionic (odd) term of the form

$$(\kappa_{AF})^\alpha \epsilon_{\alpha\beta\mu\nu}A^\beta F^{\mu\nu} ;$$

however, since this gives rise to negative contributions to the canonical energy and may lead to instabilities in the theory [22.20, 21], it is usually considered to vanish, $k_{AF} = 0$ – which is experimentally supported.

Given the suggestive notation above, one naturally obtains a generalized Dirac equation

$$(\mathrm{i}\Gamma^\mu\partial_\mu - M)\psi = 0 , \tag{22.12}$$

together with generalized inhomogeneous Maxwell equations (without sources)

$$\partial_\nu F^{\mu\nu} + (\kappa_F)^\mu{}_{\nu\alpha\beta}\partial^\nu F^{\alpha\beta} = 0 , \tag{22.13}$$

while the homogeneous Maxwell equations remain the same. The full set may be suggestively recast into the usual counterpart, $\partial_\mu F^{\mu\nu} = 0$, but with the deformed constitutive relations for the medium

$$\begin{pmatrix} \boldsymbol{D} \\ \boldsymbol{H} \end{pmatrix} = \begin{pmatrix} \epsilon_0(\tilde{\epsilon}_r + \kappa_{DE}) & \sqrt{\frac{\epsilon_0}{\mu_0}}\kappa_{DB} \\ \sqrt{\frac{\epsilon_0}{\mu_0}}\kappa_{HE} & \mu_0^{-1}(\tilde{\mu}_r^{-1} + \kappa_{HB}) \end{pmatrix} \begin{pmatrix} \boldsymbol{E} \\ \boldsymbol{B} \end{pmatrix} , \tag{22.14}$$

where $\tilde{\epsilon}_r$ and $\tilde{\mu}_r$ are the electric permittivity and magnetic permeability matrices, respectively (proportional to the 3×3 identity matrix for linear, homogeneous and isotropic mediums), and one defines

$$\begin{aligned} {\kappa_{DE}}^{ij} &= -2{\kappa_F}^{0i0j} , \\ {\kappa_{HB}}^{ij} &= \frac{1}{2}\epsilon^{ikl}\epsilon^{jmn}{\kappa_F}^{klmn} , \\ {\kappa_{DB}}^{ij} &= -{\kappa_{HE}}^{ji} = {\kappa_F}^{0ikl}\epsilon^{jkl} . \end{aligned} \tag{22.15}$$

Hence, one has the ingredients to perform a thorough analysis of a possible breaking of Lorentz invariance involving charged particles and light.

In four spacetime dimensions, renormalizability of standard model operators requires that these have a mass dimension $d \leq 4$; however, in principle, Lorentz breaking operators with any mass dimension d could also appear in the Lagrangian of the effective field theory extending the standard model at low energies.

If the lower-dimensional operators $d \leq 4$ are not adequately suppressed at low-energy scales, they dominate the higher-dimensional ones and lead to unacceptably high corrections to the deformed dispersion relation and have the form $f^{(n)} p^n M_P^{2-n}$, with $f^{(n)} \approx 1$. Moreover, radiative corrections lead to additional linear and quadratic terms of the form $M_P p + f^{(n)} p^2$.

Since it is known experimentally that the dispersion relation of SR holds with great accuracy (Sect. 22.3.3), one requires an unnatural fine-tuning of the dimensionless coefficients affecting these operators [22.22], so that additional linear and quadratic terms in the deformed dispersion relation cancel out. An explicit computation of the dispersion relation from the mSME for fermions can be found in [22.23].

As it turns out, one may resort to partial discrete symmetries that remain after the main one is broken: a natural candidate is *CPT*, as the odd Lorentz-invariant operators of the *mSME* are restricted (and thus made compatible with the experimental bounds) if one enforces this symmetry in the theory. Even operators may also be suppressed if one invokes supersymmetry as a natural invariance of Nature, although consistency requires that allowed Lorentz symmetry breaking operators involving supersymmetric partners are also considered, and even operators remain dangerously unrestricted.

The kinetic and dynamic formalisms presented above are all naturally intertwined, and may be correlated although the underlying physical theory remains unknown – that is, a particular Lorentz breaking contribution to the effective field theory envisaged in the dynamical framework naturally translates into specific modified dispersion relations [22.24, 25], while phenomenological terms considered in the kinetic approach can, in principle, be traced back to relevant operators at the low-energy level [22.17].

22.3 Testing General Relativity

Having discussed above how the foundational principles of SR can be tested, one now focuses on GR and its current experimental status. The first experimental confirmation of GR appeared in 1915, when it successfully accounted for the discrepancy with the Newtonian estimate for the advance of the perihelion precession of Mercury's orbit with no adjustable parameters. Shortly after, the famous 1919 expedition by Eddington produced observations of stellar lines-of-sight during a solar eclipse that confirmed another prediction of GR, namely that the deflection angles due to light bending around the gravitational field of the Sun should be twice the value obtained from Newtonian and EP arguments. This propelled GR into notoriety and turned its creator into the first scientific star of the world.

Since then, GR has been extensively tested in the Solar System, with all data obtained so far being consistent with its predictions. As time has gone by, these tests have grown more and more precise: from the ≈ 0.2 accuracy of microwave ranging to the Viking Lander on Mars in 1976 [22.26–28] and 0.15 accuracy of spacecraft and planetary radar observations [22.29] to an order of magnitude gain via the astrometric observations of quasars on the solar background performed with very-long baseline interferometry [22.30–32] and lunar laser ranging precision measurements of the lunar orbit (with accuracies of ≈ 0.045 and ≈ 0.011, respectively) [22.33–39]. This was pushed even further by the 2003 experiments with the Cassini spacecraft, which improved the testing accuracy down to ≈ 0.0023 [22.40].

Observations of binary millisecond pulsars lend further support for GR; indeed, the physical processes occurring in the strong gravitational field regime within these relativistic object are of considerable interest, given the possibility of testing relativity in a distinct dynamical environment. Pulsar tests of strong-field gravity were first formulated in [22.41], with initial tests being performed with PSR1534 [22.42]. Strong-field gravitational tests and their theoretical rationale were examined in [22.43–45]. Pulsar data were recently analyzed to test GR to ≈ 0.04 at a 3σ confidence level [22.46].

22.3.1 Metric Theories of Gravity and PPN Formalism

In this section, we present the formalism used to interpret observations in weak field and slow motion

approximation, conditions found in the Solar System; this formalism provides a rigorous framework to study increasingly accurate experiments and to establish stringent constraints on deviations from GR and its fundamental tenets.

The distribution of matter in this approximation is commonly represented by a perfect fluid model [22.47–50] with an energy-momentum tensor $\widehat{T}^{mn}$ given by

$$\widehat{T}^{mn} = \sqrt{-g}\left([\rho_0(1+\Pi)+p]\,u^m u^n - p g^{mn}\right), \tag{22.16}$$

where ρ_0 is the mass density of the ideal fluid in coordinates of the comoving frame of reference, u^k are the components of invariant four-velocity of a fluid element, and $p(\rho)$ is the isentropic pressure connected with the energy density by an equation of state $p = p(\rho)$. The quantity $\rho\Pi$ is the density of internal energy of an ideal fluid; the definition of Π arises from the first law of thermodynamics, according to the equation $u^n\left(\Pi_{;n} + p\,(1/\widehat{\rho})_{;n}\right) = 0$, where $\widehat{\rho} = \sqrt{-g}\rho_0 u^0$ is the conserved mass density [22.1, 49–51]. Considering the energy–momentum tensor, the solutions of the gravitational field equations for a given theory of gravity can be found.

An alternative methodology, valid for both weak and strong regimes of GR and an arbitrary energy–stress tensor, builds upon a *Maxwell-like* expansion of the metric and the Blanchet–Damour multipole framework [22.52–56]; the study of a general N-body problem in a weak-field and slow motion approximation was developed in [22.57].

Despite the widely different principles underlying metric theories of gravity, they all share the feature that the gravitational field directly affects the matter through the metric tensor g_{mn}, which is determined from the field equations. Thus, the metric expresses the properties of a particular gravitational theory and carries information about the bodies' gravitational field – contrasting with the flat metric of Newtonian gravity and its interpretation in terms of forces acting at a distance.

The so-called PPN formalism generalizes the phenomenological parameterization of the gravitational metric tensor field first discussed by Eddington in a limited context [22.58–60]. This method assumes slowly moving bodies and weak inter-body gravity, and is valid for a broad class of metric theories. The PPN parameters that appear in the expansion of the metric characterize each theory of gravity and are individually associated with the underlying symmetries and laws of invariance. If, for simplicity, one assumes Lorentz and local position invariance and conservation of total momentum conservation, the metric tensor in four dimensions in the PPN-gauge is given by

$$\begin{aligned}
g_{00} &= -1 + 2U - 2\beta U^2 - 2\xi\Phi_W \\
&\quad + (2\gamma + 2 + \alpha_3 + \zeta_1 - 2\xi)\Phi_1 \\
&\quad + 2(3\gamma - 2\beta + 1 + \zeta_2 + \xi)\Phi_2 + 2(1+\zeta_3)\Phi_3 \\
&\quad + 2(3\gamma + 3\zeta_4 - 2\xi)\Phi_4 - (\zeta_1 - 2\xi)\mathcal{A} \\
&\quad - (\alpha_1 - \alpha_2 - \alpha_3)w^2 U - \alpha_2 w^i w^j U_{ij} \\
&\quad + (2\alpha_3 - \alpha_1)w^i V_i + O(\epsilon^3), \\
g_{0i} &= -\tfrac{1}{2}(4\gamma + 3 + \alpha_1 - \alpha_2 + \zeta_1 - 2\xi)V_i \\
&\quad - \tfrac{1}{2}(1 + \alpha_2 - \zeta_1 + 2\xi)W_i \\
&\quad - \tfrac{1}{2}(\alpha_1 - 2\alpha_2)w^i U - \alpha_2 w^j U_{ij} + O(\epsilon^{5/2}), \\
g_{ij} &= (1 + 2\gamma U)\delta_{ij} + O(\epsilon^2),
\end{aligned} \tag{22.17}$$

setting $\hbar = c = G = 1$ and using the metric signature convention $(- + ++)$.

The order of magnitude of the various terms is determined according to the estimates $U \approx v^2 \approx \Pi \approx p/\rho \approx \epsilon$, $v^i \approx |\mathrm{d}/\mathrm{d}t|/|\mathrm{d}/\mathrm{d}x| \approx \epsilon^{1/2}$, and all possible potentials are considered up to the desired post-Newtonian order. Considering (22.16), these generalized gravitational potentials, of the same order as U^2,

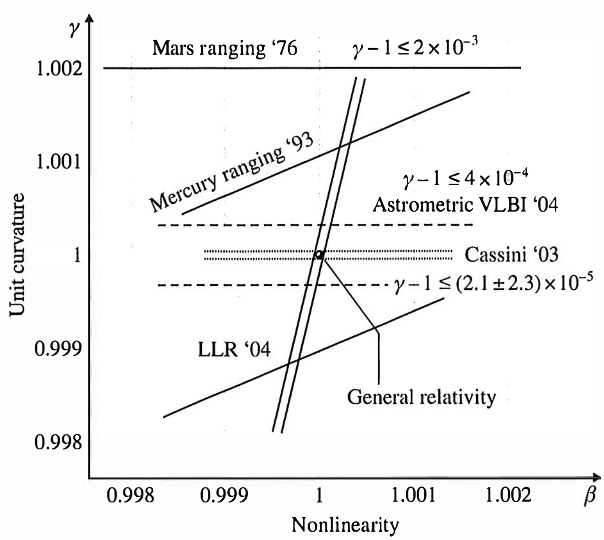


Fig. 22.1 The progress in determining the PPN parameters γ and β for the last 30 y (After [22.61])

are given by

$$
\begin{aligned}
U &= \int \frac{\rho'}{|x-x'|} \mathrm{d}^3x' , \\
U_{ij} &= \int \frac{\rho'(x-x')_i(x-x')_j}{|x-x'|^3} \mathrm{d}^3x' , \\
\Phi_W &= \int \frac{\rho'\rho''(x-x')}{|x-x'|^3} \\
&\quad \times \left(\frac{x'-x''}{|x-x''|} - \frac{x-x''}{|x'-x''|}\right) \mathrm{d}^3x' \mathrm{d}^3x'' , \\
\mathcal{A} &= \int \frac{\rho'[v'\cdot(x-x')]^2}{|x-x'|^3} \mathrm{d}^3x' , \\
\Phi_1 &= \int \frac{\rho' v'^2}{|x-x'|} \mathrm{d}^3x' , \\
\Phi_2 &= \int \frac{\rho' U'}{|x-x'|} \mathrm{d}^3x' , \\
\Phi_3 &= \int \frac{\rho' \Pi'}{|x-x'|} \mathrm{d}^3x' , \\
\Phi_4 &= \int \frac{p'}{|x-x'|} \mathrm{d}^3x' , \\
V_i &= \int \frac{\rho' v'_i}{|x-x'|} \mathrm{d}^3x' , \\
W_i &= \int \frac{\rho'[v'\cdot(x-x')](x-x')_i}{|x-x'|^3} \mathrm{d}^3x' .
\end{aligned}
\tag{22.18}
$$

A particular metric theory of gravity in the PPN formalism is fully characterized by means of the 11 PPN parameters shown in (22.17) [22.1, 62]: these have clear physical meaning and concern a particular symmetry, conservation law, or fundamental tenet of the structure of spacetime: the parameter β is the measure of the nonlinearity of the law of superposition of the gravitational fields (or its metricity) in a theory of gravity, while γ represents the measure of the curvature of the spacetime created per unit rest mass; the group of parameters $\alpha_1, \alpha_2, \alpha_3$ quantify the violation of Lorentz invariance (i. e., the existence of the privileged reference frame), the parameter ζ quantifies the violation of local position invariance, and the parameters $\alpha_3, \zeta_1, \zeta_2, \zeta_3, \zeta_4$ indicate a possible violation of the conservation of total momentum.

Since GR satisfies all of the above principles, it is naturally signaled by the vanishing of all PPN parameters except β and γ. Brans–Dicke theory [22.63], perhaps the best known of the alternative theories of gravity, endowed with an additional scalar field and arbitrary coupling constant ω, yields a decreasing spacetime curvature per unit rest mass, while preserving the remaining symmetries: its nonvanishing PPN parameters are thus $\beta = 1, \gamma = (1+\omega)/(2+\omega)$. More general scalar tensor theories yield values of β different from unity [22.64].

The PPN metric tensor, given by (22.17)–(22.18), is used to generate the equations of motion for the bodies under scrutiny (planets, satellites, etc.), which are translated into orbit determination numerical codes [22.62, 65–67], as well as used in the analysis of gravitational experiments in the Solar System [22.1, 61]. Table 22.1 and Fig. 22.1 show the latest bounds on the Eddington parameters β and γ and the history of increasingly accurate experiments.

The foundations of GR and the current experimental verification of their validity are now discussed. For this, one recalls its basic tenets:

1. Weak equivalence principle (WEP) (also known as the principle of universality of the free fall): freely falling bodies have the same acceleration in the same gravitational field, independently of their compositions (see Sect. 22.3.2).
2. Local Lorentz invariance (LLI): the rate of clocks is independent of the velocity of the clock (Sect. 22.3.3).
3. Local position invariance (LPI): the rate of clocks is independent of the spacetime position of the clock (see Sect. 22.3.4).

22.3.2 The Equivalence Principle (EP)

Almost every theory of gravitation has addressed the issue concerning the equivalence between inertial and passive gravitational mass, starting with Newton himself. Almost one century ago, Einstein followed through by declaring that all nongravitational laws should behave in free-falling frames as if gravity were absent. This postulate implies that identical accelerations should be experienced by objects with different compositions in the same gravitational field – so that gravity becomes a geometrical property of spacetime, as posited by GR. As it turns out, this EP can be cast in both a weak and a strong version, as addressed below.

The Weak Equivalence Principle (WEP)

The weak form of EP states that the gravitational properties of all interactions except gravity obey the EP. The concerned *charges* are the nuclear-binding energy differences between test masses, their neutron-to-proton ratios or atomic charges, amongst others. The

Table 22.1 Accuracy of determination of the PPN parameters γ and β [22.2, 39, 61]

PPN parameter	Experiment	Result
$\gamma - 1$	Cassini 2003 spacecraft radio-tracking	2.3×10^{-5}
	Observations of quasars with astrometric VLBI	3×10^{-4}
$\beta - 1$	Helioseismology bound on perihelion shift	3×10^{-3}
	LLR test of the SEP, assumed: $\eta = 4\beta - \gamma - 3$ and the Cassini result for PPN γ	1.1×10^{-4}

equivalence between gravitational and inertial masses implies that distinct neutral massive test bodies have the same free fall acceleration in an external gravitational field [22.68], with the latter inducing only a tidal force [22.69].

According to GR, the spacetime curvature caused by a massive body scatters light rays passing in its vicinity achromatically. The Sun is the dominating contributor to this effect in the Solar System, deflecting the light by as much as $1.75'' \cdot (R_\odot/b)$, where $R_\odot$ is the solar radius and b is the impact parameter. In 1919, the famous Eddington expedition confirmed that photons fall freely according to the predictions of GR; although the original experiment had only a 10% accuracy, the light bending measured in a solar conjunction by the Cassini spacecraft has improved this type of measurement to the current figure of 0.0023% [22.40].

WEP also implies a Doppler frequency shift $\Delta\nu$ induced on light by the variation of the gravitational potential. This was confirmed in 1960 by the eponymous Pound–Rebka experiment, which produced

$$\frac{\Delta\nu}{\nu} = \frac{gH}{c^2} = (2.57 \pm 0.26) \times 10^{-15}\,, \qquad (22.19)$$

where g is the acceleration of gravity and H the height of the fall [22.70, 71].

Notwithstanding some formidable experimental obstacles, the free fall of antiprotons and antihydrogen (or other antiparticles) could provide yet another test of the WEP (see [22.72] for a thorough review). This would allow one to probe as to what extent gravity respects the CPT symmetry of local quantum field theories – specifically, if antiparticles fall as particles in a gravitational field. The ATHENA (Apparatus for High Precision Experiments on Neutral Antimatter) and the ATRAP (antihydrogen trap) collaborations at CERN (European Organization for Nuclear Research) have developed the capability of storing antiprotons and creating an antihydrogen atom [22.73, 74], but no experiment along these lines has been performed so far.

A test of WEP involving neutral kaons was performed by the CPLEAR ring (charge parity low-energy antiproton ring) collaboration [22.75], producing limits of 6.5, 4.3 and 1.8×10^{-9}, respectively, for scalar, vector, and tensor potentials originating from the Sun with a range much greater than 1 AU acting on kaons and antikaons. These relevant results do not probe possible baryon number-dependent interactions, and are thus complementary to the desirable antiprotons and antihydrogen atom experiments mentioned above.

Most metric theories of gravitation inherently uphold WEP, although some predict additional forces that lead to composition-dependent deviations from geodesic motion (e.g., if a nonminimal coupling between matter and curvature is present [22.76, 77]). Similarly, almost all extensions to the standard model of particle physics predict new forces that induce apparent violations of EP [22.78, 79]; this is most apparent if macroscopic-range fields are present, so that exchange forces that couple to generalized charges arise, rather than just to mass/energy as with gravity [22.80, 81].

Laboratory tests of WEP can be made by comparing the free fall accelerations a_1 and a_2 of different test bodies. If these are at the same distance from the source of the external gravitational field, the breaking of the WEP is elegantly gauged through the quantity

$$\begin{aligned}\frac{\Delta a}{a} = \frac{2(a_1 - a_2)}{a_1 + a_2} &= \left(\frac{M_G}{M_I}\right)_1 - \left(\frac{M_G}{M_I}\right)_2 \\ &= \Delta\left(\frac{M_G}{M_I}\right)\,,\end{aligned} \qquad (22.20)$$

where M_G and M_I are, respectively, the gravitational and inertial masses of each body.

Other tests conducted so far have validated WEP for elementary particles. For the neutron, an interferometry experiment showed that a neutron beam split by a silicon crystal and traveling through different gravitational paths interferes as predicted by quantum mechanics, with a gravitational potential given by Newtonian gravity – providing a striking confirmation of WEP using an elementary hadron [22.82]. Since then, gravitational atom interferometric measurements have probed WEP down to a precision of 3×10^{-8} [22.83].

The ratio of gravitational to inertial masses of test bodies has been determined, with an upper limit for $|1 - M_G/M_I|$ of $\approx 10^{-11}$ in 1964 [22.84], $\approx 10^{-12}$ in 1972 (reconfirmed in 1994) [22.85, 86] and, more recently, 1.4×10^{-13} [22.87] (see [22.88] for a review). These increasingly precise experiments further show that strong, weak, and electromagnetic interactions contribute equally to the passive gravitational and inertial masses of test bodies.

One decade ago, gravitational bound states of neutrons were confirmed by *Nesvizhevsky* and collaborators [22.89, 90], who set up a realization of a conceptual experiment proposed in 1978 [22.91]. In this experiment, ultracold neutrons from a source at the Institute Laue–Langevin reactor in Grenoble fall under the influence of the Earth's gravitational field towards a horizontal mirror, with a minimum measurable energy of 1.4×10^{-12} eV corresponding to a vertical velocity of 1.7 cm/s (a more intense beam and an enclosure mirrored on all sides could lower the latter by six orders of magnitude). The neutrons were found not to fall continuously; rather, they jumped between different vertical levels, as predicted by quantum mechanics.

Improved experiments probing gravity through quantum systems clearly open the possibility of testing novel concepts related to the unification of GR and quantum mechanics (in the low-energy regime), such as noncommutative formulations of the latter [22.92] – as well as detecting the transition between the classical and quantum description of a system as a function of its dimensions [22.93].

An analysis of the lunar laser ranging data showed the absence of any composition-dependent acceleration effects [22.94]. In astronomical measurements, one should consider the gravitational self-energy contributions to the inertial and gravitational masses of the bodies [22.58], whereas these are negligible in test masses used in laboratory environments. Considering the gravitational self-energy leads one to scrutinize the strong equivalence principle, as is discussed below.

The Strong Equivalence Principle (SEP)

The strong formulation of EP addresses the gravitational behavior arising from gravitational energy itself, thus expressing the nonlinearity of gravitation. It states that not only the outcome of gravitational experiments, but indeed any measurement concerning other interactions, are independent of the velocity and position of the laboratory. Being an integral part of EP, SEP is enforced by GR. However, many theories of gravity do not respect this assumption: for instance, scalar-tensor theories typically violate SEP [22.33, 58, 95, 96], e.g., by positing different couplings between these fields and different species of matter. This leads not only to a difference in free fall and related tests, but also in nongravitational experiments.

The fractional contributions to the mass by the gravitational self-energy of a body is the most relevant quantity for probing the validity of SEP. The previously described PPN formalism is particularly suited to the description of astronomical tests; using it, one may cast this quantity as

$$\Delta \left(\frac{M_G}{M_I} \right)_{SEP} = \eta \left(\frac{\Omega}{Mc^2} \right) , \tag{22.21}$$

where Mc^2 is the total mass–energy of the test body, Ω its negative gravitational self-energy, and η a dimensionless constant for SEP violation [22.33, 58, 95]. It is expressed by a combination of PPN parameters, so that in fully-conservative, Lorentz-invariant theories of gravity [22.1, 2], it reads $\eta = 4\beta - \gamma - 3$ (so that the values $\beta = \gamma = 1$ characterizing GR yield $\eta = 0$).

The self-energy of a body B is given by

$$\left(\frac{\Omega}{Mc^2} \right)_B = -\frac{G}{2M_B c^2} \int_B \mathrm{d}^3 x\, \mathrm{d}^3 y \frac{\rho(x)\rho(y)}{|x - y|} . \tag{22.22}$$

A sphere with a radius R and uniform density has $\Omega/Mc^2 = -3GM/5Rc^2 = -(3/10)(v_E/c)^2$, where v_E is its escape velocity. A more realistic value may be obtained by numerically integrating the expression above using its known structural features; in the case of the Sun, this yields $(\Omega/Mc^2)_\odot \approx -3.52 \times 10^{-6}$ [22.97], which should be compared with the typical magnitude $\approx 10^{-25}$ for laboratory-sized bodies. Thus, while an experimental accuracy of 10^{-13} [22.87] is sufficient to significantly constrain violations of WEP, it does not allow for a stringent test of SEP – hence the need for planetary-sized extended bodies, where the ratio (22.22) is much larger.

Several Solar System experiments have been suggested in order to probe the validity of SEP [22.33, 58, 98], from lunar measurements to the study of the motion of Trojan asteroids (performed more than two decades ago [22.99, 100]) or the analysis of binary pulsar data [22.101] – which takes advantage of a strong (self-)gravity regime [22.43, 44], albeit no sufficiently accurate measurements are yet available [22.102, 103]. Interplanetary spacecraft provide yet another testbed for SEP [22.68, 104].

So far, the most competitive assessment of the validity of SEP stems from the Earth-Moon-Sun system, through the analysis of lunar laser ranging (LLR) data [22.39] yielding $\Delta(M_G/M_I)_{SEP} = (-2.0 \pm 2.0) \times 10^{-13}$ (from a general breaking of EP of $\Delta(M_G/M_I)_{EP} = (-1.0 \pm 1.4) \times 10^{-13}$) – implying a SEP violation parameter $\eta = 4\beta - \gamma - 3 = (4.4 \pm 4.5) \times 10^{-4}$.

22.3.3 Local Lorentz Invariance (LLI)

Invariance under Lorentz transformations states that the laws of physics are independent of the velocity of the frame. This is the basic tenet of SR, as discussed before, and holds only locally in GR. Although current theories obey this symmetry, some results arising from string field theory hint that it may be spontaneously broken [22.105, 106], due to open string interactions and its implications at low-energy physics. If so, many implications are expected: for instance, if the contribution of Lorentz-violating interactions to the vacuum energy is approximately half of the critical density, one expects that very weak tensor-mediated interactions arise in the range $\approx 10^{-4}$ m [22.107]. Furthermore, these string interactions are the privileged contributors to the Lorentz violating terms of mSME.

The effect of our velocity relative to a putative preferred reference frame may be phenomenologically described by considering a cosmological vector field that acquires a nonvanishing minima due to a spontaneous symmetry breaking induced by a suitable potential [22.108]; a model allowing for such a spontaneous breaking of LLI has been proposed [22.109–111], leading to interesting scenarios where the inverse square law for gravity is modified by the spacetime direction chosen by the vector field [22.19].

Considerations on the dynamics of the renormalization group β-function of nonabelian gauge theories also hint that Lorentz invariance might be just a low-energy symmetry [22.112]. Lorentz violation may also induce the breaking of conformal symmetry; together with inflation, this could explain the primordial magnetic fields needed to account for the observed galactic magnetic field [22.113]. A modified gravity-induced wave dispersion derived from a violation of Lorentz invariance could be probed by astrophysical observations of distant sources of gamma radiation [22.114, 115].

A violation of this fundamental symmetry of GR is also possible with noncommutative field theories [22.116], although it may hold (at least) at first nontrivial order in perturbation theory of the noncommutative parameter [22.117–121]. Other theories that may entail a breaking of Lorentz invariance include loop quantum gravity [22.122, 123], spacetime foam scenarios [22.124, 125], and models exhibiting a spacetime variation of fundamental coupling constants [22.126, 127] (see [22.128] for a review of high-energy Lorentz symmetry breaking).

A violation of Lorentz invariance could break the fundamental CPT symmetry of local quantum field theories [22.129, 130] – a prospect that can be tested in neutral-meson [22.131, 132] experiments, Penning-trap measurements [22.133, 134], and hydrogen-antihydrogen spectroscopy [22.135]. This CPT breaking could also be induced by nonlinearities in quantum mechanics, perhaps stemming from a quantum theory of gravity; the latter possibility has been probed by the CPLEAR Collaboration [22.136]. Whatever the cause, the spontaneous breaking of CPT symmetry provides, along with the violation of the baryon number, an interesting mechanism for the generation of the observed baryon asymmetry in the Universe: after the CPT and baryon number symmetries are broken in the early Universe, tensor–fermion interactions arising from string field theory give rise to a chemical potential that creates a baryon-antibaryon asymmetry in equilibrium [22.137].

Modifications of the Michelson–Morley experiment using laser interferometry are very useful for testing Lorentz symmetry breaking, by comparing the velocity of light and the maximum attainable velocity of massive particles, c_i – with a current experimental constraint of $|c^2/c_i^2 - 1| < 10^{-9}$ [22.138] (see Sect. 22.2).

The more accurate Hughes–Drever experiment probes a possible time dependence of the quadrupole splitting of nuclear Zeeman levels along Earth's orbit [22.139, 140], yielding an impressive limit of $|c^2/c_i^2 - 1| < 3 \times 10^{-22}$ [22.141] – with a follow-up study showing that a gain of up to eight orders of magnitude in accuracy is possible [22.142].

As stated before, astronomical tests are best analyzed through the use of the PPN formalism, with the α_3 parameter being related to violation of momentum conservation and the existence of a preferred reference frame ($\alpha_3 = 0$ in GR). The study of (millisecond) pulsars yields the extremely accurate limit $|\alpha_3| < 2.2 \times 10^{-20}$ [22.2, 143, 144].

An analysis of the interaction between the most energetic cosmic-ray particles and the photons from the cosmic microwave background radiation has shown that the propagation of ultra-high-energy nucleons is limited by inelastic collisions with the latter, preventing particles with energies above 5×10^{19} eV from reaching Earth

from beyond 50–100 Mpc – the so-called Greisen–Zatsepin–Kuzmin (GZK) cut-off [22.145, 146]. Events where the cosmic primaries have an estimated energy above the GZK cut-off were observed by different collaborations [22.147–153]; although the HI-RES (high resolution fly's eye) [22.154] and Auger [22.155] collaborations results have been interpreted as being consistent with the validity of this cutoff, and hence of Lorentz symmetry.

Processes such as the resonant scattering reaction $p + \gamma_{2.73K} \rightarrow \Delta_{1232}$ have been shown to be suppressed by energy-dependent effects arising from small violations of Lorentz invariance [22.156–159]. This can be used to analyze the putative existence of events above the GZK cutoff, yielding the strongest constraint of $|c^2/c_i^2 - 1| < 1.7 \times 10^{-25}$ [22.23, 160, 161].

22.3.4 Local Position Invariance (LPI)

A violation of LPI indicates that the rates of a free falling clock and one on the surface of the Earth should differ. As the WEP and LLI principles of GR benefit from the stringent bounds addressed before, experiments on the universality of the gravitational red-shift primarily probe the validity of the LPI. This may be quantified by the parameter μ measuring the deviation in the relative shift in frequency $\Delta\nu/\nu = (1+\mu)U/c^2$ when compared with GR (where $\mu = 0$).

The already discussed Pound–Rebka experiment (22.19) yields $\mu \simeq 10^{-2}$. An accurate verification of LPI was achieved through the comparison between hydrogen-maser frequencies on Earth and of a rocket flying to altitude of 10 000 km [22.162], leading to $|\mu| < 2 \times 10^{-4}$. Further considerations allow for an improvement by two orders of magnitude, $\mu < (0.1 \pm 1.4) \times 10^{-6}$ [22.163].

22.3.5 The Pioneer and Flyby Anomalies

Although not quite a direct test of SR or GR per se, the Pioneer and flyby anomalies have arisen in the literature as phenomena that, at least at a first look, challenged the common wisdom about gravity. These unaccounted behaviors of spacecraft, derived from the analyses of tracking data, have led many theoreticians to the drawing board, with suggestions that these anomalies embody new physical phenomena that could encompass a putative breaking of the basic tenets of relativity.

The Pioneer anomaly stood out as an open question in physics for more than a decade; its existence was first discussed in 1998 [22.164], when a JPL team showed that the deep tracking of the Pioneer 10 and 11 probes disagreed with the predictions of a detailed orbital determination model including GR and all relevant effects and ephemerides – but was statistically consistent with a fit to the latter plus a constant Sun-bound acceleration $a_P = (8.74 \pm 1.33) \times 10^{-10}\,\mathrm{m/s^2}$ [22.165].

This anomalous behavior was independently confirmed through alternative data analyses [22.166–168], with the first pair of studies allowing for a decreasing acceleration, instead of a constant one. Indeed, 10 years ago it was pointed out that this was compatible with an exponentially decreasing acceleration with a timescale compatible with the decay rate of the plutonium present in the radiothermal generators (RTG) and powering the spacecraft. Nonetheless, and despite studies pointing at a conventional origin for the Pioneer anomaly [22.169, 170], more specifically onboard thermal effects, this possibility was strongly rejected by the JPL team and explanations resorting to new physics appeared (see [22.76, 171–173] and references therein). It was also shown that the most considered models for the mass distribution of the Kuiper belt could not cause the anomalous acceleration [22.174] (see also [22.175]).

It was only in 2008 that a clear numerical indication that the Pioneer anomaly was of thermal origin appeared, with the radiation emitted from the RTG and the main compartment providing the additional, decaying thrust that deviated the twin probes from its predicted trajectories [22.176]. This possibility gained strength with the following independent studies [22.177–180], culminating in a recent study showing that the observed anomaly falls squarely into the predictions yielded by a model that also considers the reflection of the radiation on the parabolic dish of the high-gain antenna [22.181] – a result confirmed by a subsequent study by other teams [22.182, 183].

Thus, the Pioneer anomaly is no more and now serves as a cautionary tale against the dangers of extrapolating poorly understood conventional effects as revolutionary evidence of deviations from SR and GR. With this is mind, the more recent flyby anomaly is viewed with added scepticism, although it has so far defied any conventional explanation.

The Flyby Anomaly

The flyby anomaly is an unexpected velocity change disclosed by the analysis of several Earth gravity-assist maneuvers of the Galileo, NEAR, Cassini, and Rosetta spacecraft [22.184–186]. Following flybys of the Galileo and Rosetta missions raised some expectation of obtaining a confirmation of this phenomenon.

Table 22.2 Summary of orbital parameters of the considered Earth flybys

Mission	Date	e	Perigee (km)	v_∞ (km/s)	Δv_∞ (mm/s)	$\Delta v_\infty/v_\infty$ (10^{-6})
Galileo	1990	2.47	959.9	8.949	3.92 ± 0.08	0.438
Galileo	1992	3.32	303.1	8.877	≈ 0	−0.518
NEAR	1998	1.81	538.8	6.851	13.46 ± 0.13	1.96
Cassini	1999	5.8	1173	16.01	-2 ± 1	−0.125
Rosetta	2005	1.327	1954	3.863	1.80 ± 0.05	0.466
MESSENGER	2005	1.360	2347	4.056	0.02 ± 0.01	0.0049
Rosetta	2007	3.562	5322	9.36	≈ 0	–
Rosetta	2009	2.956	2483	9.38	≈ 0	–

However, these events yielded no further evidence of such a flyby anomaly (see Table 22.2) – in the case of the second Galileo flyby, due to the high uncertainty of the atmospheric drag, enhanced due to the very low perigee altitude of ≈ 300 km.

With the exception of the Cassini spacecraft, the involved spacecraft had no deep space network tracking during perigee passage, leading to an approximate 4 hr gap. The 10 s sampling interval for the remaining period produced a very coarse-grained distribution of data points, disabling an accurate characterization of the effect in terms of an additional force affecting the spacecraft. Thus, the flyby anomaly is signaled by the inability to fit a single hyperbolic arc to the whole flyby maneuver: two distinct *incoming* and *outgoing* arcs must be considered, with the small difference between them being interpreted as an additional boost Δv at perigee.

Despite the difficulty to assign a well-defined value, an averaged acceleration of the order of $a_F \approx 10^{-4}\,\text{m/s}^2$ may be used as a figure of merit for the flyby anomaly [22.185]. This figure allows for a comparison with several possible causes: Earth's oblateness, other Solar System bodies, relativistic corrections, atmospheric drag, Earth's albedo and infrared emissions, ocean or solid tides, solar pressure, spacecraft charging, magnetic moments, solar wind, spin-rotation coupling [22.185, 187], dark matter [22.186], etc. (Table 22.3).

Table 22.3 List of orders of magnitude of possible error sources during Earth flybys

Effect	Order of Magnitude (m/s²)
Earth's oblateness	10^{-2}
Other Solar System bodies	10^{-5}
Relativistic effects	10^{-7}
Atmospheric drag	10^{-7}
Ocean and Earth tides	10^{-7}
Solar pressure	10^{-7}
Earth's infrared emissions	10^{-7}
Spacecraft charge	10^{-8}
Earth's albedo	10^{-9}
Solar wind	10^{-9}
Magnetic moment	10^{-15}

Clearly, all these effects are much smaller than the considered value for a_F, with the exception of Earth's oblateness. However, accurate knowledge of the gravitational model of the Earth means that the origin of the flyby anomaly cannot be due to some minor deviation in the latter [22.185].

The empirical formula proposed in [22.184] is perhaps the most prominent attempt to account for the reported flyby anomalies; it proposes that the variation in magnitude and direction of the anomalous velocity change reflects the declinations of the incoming and outgoing asymptotic velocity vectors, δ_i and δ_o, respectively,

$$\frac{\Delta V_\infty}{V_\infty} = \frac{2\omega_\text{E} R_\text{E}}{c}(\cos\delta_\text{i} - \cos\delta_\text{o}) \,, \tag{22.23}$$

where ω_E is the Earth's rotation velocity and R_E its radius. This identification is suggestive, given its similarity with the term present in the outer metric due to a rotating body [22.188]

$$\begin{aligned} \mathrm{d}s^2 = &\left(1 + 2\frac{V - \Phi_0}{c^2}\right)(c\,\mathrm{d}t)^2 \\ &- \left(1 - 2\frac{V}{c^2}\right)(\mathrm{d}r^2 + r^2\mathrm{d}\Omega^2) \,, \end{aligned} \tag{22.24}$$

with

$$\frac{\Phi_0}{c^2} = \frac{V_0}{c^2} - \frac{1}{2}\left(\frac{\omega_\text{E} R_\text{E}}{c}\right)^2 , \tag{22.25}$$

where $\mathrm{d}\Omega^2 = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$ and V_0 is the Newtonian potential $V(r)$ at the equator.

However, any reasoning attempting to derive (22.23) from GR is faulty, as all relativistic effects (embodied in the above metric) have been calculated to be much lower than the typical order of magnitude a_F of the flyby anomaly – namely those induced by the rotation of the Earth: the de Sitter precession effect and frame dragging.

Furthermore, the application of (22.23) to the subsequent two flybys by the Rosetta probe in 2007 and 2008 predicted an anomalous increase in V_∞ of, respectively, 0.98 and 1.09 mm/s [22.189], but the analysis of the tracking data was not consistent with any flyby anomaly whatsoever.

Similarly to what occurred with the Pioneer anomaly, a conventional explanation for the flyby anomaly should not be dismissed. Indeed, some yet unmodeled aspect of the affected spacecrafts could lead to the observed anomalous Δv; if this is the case, the widely different designs and gravitational assists of the studied spacecrafts would naturally lead to the variations of the latter.

The opposite possibility might be more enticing, namely that the flyby anomaly is the signature of new or *exotic* physics at play. Its confirmation as a new physical force would have implications to a wide range of phenomena such as planetary orbits, and potentially lead to deepening our understanding of gravity. However, no clear-cut fundamental motivation exists for such a short-ranged force (see [22.187, 190] for an overview of some proposed physical mechanisms).

In order to settle the issue, a clear-cut confirmation of this effect is mandatory. Given the sparse number of gravitational assists available, a recent proposal [22.190] suggested that a thorough characterization of the flyby anomaly could be achieved by studying the behavior of a spacecraft in a highly elliptic orbit, such that the velocity and altitude at perigee is similar to the values depicted in Table 22.2. Such a mission could come at a low cost and would provide the desired repetition of flybys; a detailed study of its design features (e.g., thermal modeling, atmospheric drag) would allow for a clear discrimination of competing perturbations, and the use of global navigation satellite system (GNSS) tracking would provide the required tracking accuracy.

This concept could be realized via a dedicated small or micro-satellite, or as an add-on to an existing mission [22.190]. The *STE-QUEST* mission, currently under consideration by the European Space Agency, could provide the latter, given its highly elliptic orbit and use of GNSS precise orbit determination [22.191].

22.3.6 Conclusion

As was seen in the preceding sections, all of the available constraints on the validity of the founding principles of SR and GR have so far failed to crack any faults in these century-old theories, which thus remain the standard against all competitors so far.

The available experimental data fit quite well with GR, while allowing for the existence of putative extensions, provided any new effects are small at the post-Newtonian scale [22.1]. However, despite its impressive experimental success, GR cannot be regarded as a fully satisfactory theory, given its inadequacy in what concerns issues such as the existence of singularities, the cosmological constant problem (see [22.192] and references therein) and the incompatibility with existing quantization schemes.

At the largest scales, GR is compatible with cosmological data if and only if dark matter dominates at galactic and cluster scales, while the dynamics of the accelerating expansion of the Universe is controlled by dark energy.

At a more conceptual level, it was recently suggested that gravity, and GR in particular, is an emerging property arising from more fundamental tenets such as the holographic principle and Bekenstein's entropy-energy limit [22.193].

This perspective leads to new challenges and may imply, for instance, that WEP might be violated if spacetimes admit a phase-space noncommutative geometry [22.194].

References

22.1 C.M. Will: *Theory and Experiment in Gravitational Physics* (Cambridge University Press, Cambridge 1993)

22.2 C.M. Will: The confrontation between general relativity and experiment, Living Rev. Relativ. **9**, 3 (2006)

22.3 O. Bertolami, J. Páramos, S.G. Turyshev: General theory of relativity: Will it survive the next decade? In: *Lasers, Clocks, and Drag-Free: Technologies for Future Exploration in Space and Tests of Gravity*, ed. by H. Dittus, C. Lämmerzahl, S. Turyshev (Springer, Berlin 2006), gr-qc/0602016

22.4 J.R. Lucas, P.E. Hodgson: *Spacetime and Electromagnetism: An Essay on the Philosophy of the Special Theory of Relativity* (Oxford University Press, Oxford 1990)

22.5 A. Albrecht, J. Magueijo: Time varying speed of light as a solution to cosmological puzzles, Phys. Rev. D **59**, 43516 (1999)

22.6 C. Lämmerzahl, C. Braxmaier, H. Dittus, H. Müller, A. Peters, S. Schiller: Kinematical test theories for special relativity, Int. J. Modern Phys. D **11**, 1109 (2002)

22.7 R. Mansouri, R. Sexl: A test theory of special relativity: I. Simultaneity and clock synchronization, Gen. Relativ. Gravit. **8**, 497–513 (1977)

22.8 S. Herrmann, A. Senger, K. Möhle, M. Nagel, E.V. Kovalchuk, A. Peters: Rotating optical cavity experiment testing Lorentz invariance at the 10^{-17} level, Phys. Rev. D **80**, 105011 (2009)

22.9 M.E. Tobar, P. Wolf, S. Bize, G. Santarelli, V. Flambaum: Testing local Lorentz and position invariance and variation of fundamental constants by searching the derivative of the comparison frequency between a cryogenic sapphire oscillator and hydrogen maser, Phys. Rev. D **81**, 022003 (2010)

22.10 S. Reinhardt, G. Saathoff, H. Buhr, L.A. Carlson, A. Wolf, D. Schwalm, S. Karpuk, C. Novotny, G. Huber, M. Zimmermann, R. Holzwarth, T. Udem, T.W. Hänsch, G. Gwinner: Test of relativistic time dilation with fast optical atomic clocks at different velocities, Nat. Phys. **3**, 861–864 (2007)

22.11 A.P. Lightman, D.L. Lee: Restricted proof that the weak equivalence principle implies the Einstein equivalence principle, Phys. Rev. D **8**, 364–376 (1973)

22.12 F. Dowker, J. Henson, R.D. Sorkin: Quantum gravity phenomenology, Lorentz invariance and discreteness, Modern Phys. Lett. A **19**, 1829–1840 (2004)

22.13 Y.J. Ng, H. van Dam: Limit to space-time measurement, Modern Phys. Lett. A **9**, 335 (1994)

22.14 K. Shiokawa: Mesoscopic fluctuations in stochastic spacetime, Phys. Rev. D **62**, 024002 (2000)

22.15 D. Colladay, V.A. Kostelecký: Lorentz-violating extension of the standard model, Phys. Rev. D **58**, 116002 (1998)

22.16 V.A. Kostelecký, M. Mewes: Electrodynamics with Lorentz-violating operators of arbitrary dimension, Phys. Rev. D **80**, 015020 (2009)

22.17 V.A. Kostelecký, M. Mewes: Signals for Lorentz violation in electrodynamics, Phys. Rev. D **66**, 56005 (2002)

22.18 V.A. Kostelecký: Gravity, Lorentz violation, and the standard model, Phys. Rev. D **69**, 105009 (2004)

22.19 O. Bertolami, J. Páramos: Vacuum solutions of a gravity model with vector-induced spontaneous Lorentz symmetry breaking, Phys. Rev. D **72**, 44001 (2005)

22.20 R. Jackiw, V.A. Kostelecký: Radiatively induced Lorentz and CPT violation in electrodynamics, Phys. Rev. Lett. **82**, 3572 (1999)

22.21 M. Perez-Victoria: Physical (ir)relevance of ambiguities to Lorentz and CPT violation in QED, J. High Energy Phys. **04**, 032 (2001)

22.22 J. Collins, A. Perez, D. Sudarsky, L. Urrutia, H. Vucetich: Lorentz invariance and quantum gravity: An additional fine-tuning problem?, Phys. Rev. Lett. **93**, 191301 (2004)

22.23 O. Bertolami, C.S. Carvalho: Proposed astrophysical test of Lorentz invariance, Phys. Rev. D **61**, 103002 (2000)

22.24 R.C. Myers, M. Pospelov: Ultraviolet modifications of dispersion relations in effective field theory, Phys. Rev. Lett. **90**, 211601 (2003)

22.25 O. Bertolami, J.G. Rosa: New bounds on cubic Lorentz-violating terms in the fermionic dispersion relation, Phys. Rev. D **71**, 097901 (2005)

22.26 I.I. Shapiro, C.C. Counselman, R.W. King: Verification of the principle of equivalence for massive bodies, Phys. Rev. Lett. **36**, 555–558 (1976)

22.27 R.D. Reasenberg, I.I. Shapiro, P.E. MacNeil, R.B. Goldstein, J.C. Breidenthal, J.P. Brenkle, D.L. Cain, T.M. Kaufman, T.A. Komarek, A.I. Zygielbaum: Viking relativity experiment: Verification of signal retardation by solar gravity, Astrophys. J. **234**, L219–L221 (1979)

22.28 J. Lorell, I.I. Shapiro: Mariner 9 celestial mechanics experiment: A status report, J. Geophys. Res. **78**(20), 4327–4329 (1973)

22.29 J.D. Anderson, E.L. Lau, S.G. Turyshev, J.G. Williams, M.M. Nieto: Recent results for solar-system tests of general relativity, Bull. Am. Astron. Soc. **34**, 833 (2002)

22.30 D.S. Robertson, W.E. Carter, W.H. Dillinger: New measurement of solar gravitational deflection of radio signals using VLBI, Nature **349**, 768–770 (1991)

22.31 D.E. Lebach, B.E. Corey, I.I. Shapiro, M.I. Ratner, J.C. Webber, A.E.E. Rogers, J.L. Davis, T.A. Herring: Measurement of the solar gravitational deflection of radio waves using very-long-baseline interferometry, Phys. Rev. Lett. **75**, 1439–1442 (1995)

22.32 S.S. Shapiro, J.L. Davis, D.E. Lebach, J.S. Gregory: Measurement of the solar gravitational deflection of radio waves using geodetic very-long-baseline interferometry data, 1979-1999, Phys. Rev. Lett. **92**, 121101 (2004)

22.33 K. Nordtvedt: Testing relativity with laser ranging to the moon, Phys. Rev. **170**, 1186–1187 (1968)

22.34 K. Nordtvedt: Lunar laser ranging reexamined: The non-null relativistic contribution, Phys. Rev. D **43**, 3131–3135 (1991)

22.35 K. Nordtvedt: 30 years of lunar laser ranging and the gravitational interaction, Class. Quantum Gravity **16**(12A), A101 (1999)

22.36 K. Nordtvedt: Lunar laser ranging: A comprehensive probe of post-Newtonian gravity, Villa Mondragone International School of Gravitation and Cosmology (2003), gr-qc/0301024

22.37 J.G. Williams, X.X. Newhall, J.O. Dickey: Relativity parameters determined from lunar laser ranging, Phys. Rev. D **53**, 6730–6739 (1996)

22.38 J.G. Williams, J.D. Anderson, D.H. Boggs, E.L. Lau, J.O. Dickey: Solar system tests for changing gravity, Bull. Am. Astron. Soc. **33**, 836 (2001)

22.39 J.G. Williams, S.G. Turyshev, D.H. Boggs: Progress in lunar laser ranging tests of relativistic gravity, Phys. Rev. Lett. **93**, 261101 (2004)

22.40 B. Bertotti, L. Iess, P. Tortora: A test of general relativity using radio links with the Cassini spacecraft, Nature **425**, 374 (2003)

22.41 T. Damour, J.H. Taylor: Strong field tests of relativistic gravity and binary pulsars, Phys. Rev. D **45**, 1840–1868 (1992)

22.42 J.N. Taylor, A. Wolszczan, T. Damour: Experimental constraints on strong field relativistic gravity, Nature **355**, 132 (1993)

22.43 T. Damour, G. Esposito-Farese: Testing gravity to second post-Newtonian order: A field theory approach, Phys. Rev. D **53**, 5541–5578 (1996)

22.44 T. Damour, G. Esposito-Farese: Tensor – scalar gravity and binary pulsar experiments, Phys. Rev. D **54**, 1474–1491 (1996)

22.45 T. Damour, G. Esposito-Farese: Gravitational wave versus binary – pulsar tests of strong field gravity, Phys. Rev. D **58**, 042001 (1998)

22.46 C. Lange, F. Camilo, N. Wex, M. Kramer, D.C. Backer, A.G. Lyne, O. Doroshenko: Precision timing measurements of PSR J1012+5307, Mon. Not. R. Astron. Soc. **326**, 274 (2001)

22.47 V.A. Fock: On movement of finite masses in the general theory of relativity, J. Phys. USSR **1**, 81 (1939), in Russian

22.48 V.A. Fock: Three lectures on relativity theory, Rev. Modern Phys. **29**, 325 (1957)

22.49 V.A. Fock: *The Theory of Space, Time and Gravitation* (Pergamon, Oxford 1959)

22.50 S. Chandrasekhar: The post-newtonian equations of hydrodynamics in general relativity, Astrophys. J. **142**, 1488 (1965)

22.51 V. Brumberg: *Essential Relativistic Celestial Mechanics* (Taylor Francis, Boca Raton 1991)

22.52 T. Damour, M. Soffel, C. Xu: General-relativistic celestial mechanics. I. Method and definition of reference systems, Phys. Rev. D **43**, 3273–3307 (1991)

22.53 T. Damour, M. Soffel, C. Xu: General relativistic celestial mechanics. 3. Rotational equations of motion, Phys. Rev. D **47**, 3124–3135 (1993)

22.54 T. Damour, M. Soffel, C. Xu: General relativistic celestial mechanics. 4: Theory of satellite motion, Phys. Rev. D **49**, 618–635 (1994)

22.55 L. Blanchet, T. Damour, B.R. Iyer, C.M. Will, A.G. Wiseman: Gravitational radiation damping of compact binary systems to second post-Newtonian order, Phys. Rev. Lett. **74**, 3515–3518 (1995)

22.56 T. Damour, D. Vokrouhlicky: Conservation laws for systems of extended bodies in the first post-Newtonian approximation, Phys. Rev. D **52**, 4455–4461 (1995)

22.57 S. Kopeikin, I. Vlasov: Parameterized post-Newtonian theory of reference frames, multipolar expansions and equations of motion in the *N*-body problem, Phys. Rep. **400**, 209–318 (2004)

22.58 K. Nordtvedt: Equivalence principle for massive bodies. 1. Phenomenology, Phys. Rev. **169**, 1014–1016 (1968)

22.59 C.M. Will: Theoretical frameworks for testing relativistic gravity. III. Conservation laws, Lorentz invariance, and values of the PPN parameters, Astrophys. J. **169**, 125 (1971)

22.60 C.M. Will, K. Nordtvedt: Conservation laws and preferred frames in relativistic gravity. I. Preferred-frame theories and an extended PPN formalism, Astrophys. J. **177**, 757 (1972)

22.61 S.G. Turyshev, J.G. Williams, K. Nordtvedt, M. Shao, T.W. Murphy: 35 years of testing relativistic gravity: Where do we go from here?, Lecture Notes in Physics **648**, 311–330 (2004)

22.62 S.G. Turyshev: Relativistic navigation: A Theoretical foundation (1996) gr-qc/9606063

22.63 C. Brans, R.H. Dicke: Mach's principle and a relativistic theory of gravitation, Phys. Rev. **124**, 925–935 (1961)

22.64 T. Damour, K. Nordtvedt: Tensor – scalar cosmological models and their relaxation toward general relativity, Phys. Rev. D **48**, 3436–3450 (1993)

22.65 T.D. Moyer: Transformation from proper time on Earth to coordinate time in solar system barycentric space-time frame of reference, Celest. Mech. **23**, 33 (1981)

22.66 T.D. Moyer: Transformation from proper time on earth to coordinate time in solar system barycentric space-time frame of reference – Part two, Celest. Mech. **23**, 57 (1981)

22.67 E.M. Standish, X.X. Newhall, J.G. Williams, D.K. Yeomans: Orbital ephemerides of the Sun, Moon, and planets. In: *Expalnatory Supplement to the Astronomical Almanac*, ed. by P.K. Seidelmann (University Science Books Mill Valley, Washington D.C. 1992)

22.68 J.D. Anderson, M. Gross, K. Nordtvedt, S.G. Turyshev: The Solar test of the equivalence principle, Astrophys. J. **459**, 365–370 (1996)

22.69 J.L. Singe: *Relativity: The General Theory* (North-Holland, Amsterdam 1960)

22.70 R.V. Pound, G.A. Rebka Jr.: Gravitational red-shift in nuclear resonance, Phys. Rev. Lett. **3**, 439–441 (1959)

22.71 R.V. Pound, G.A. Rebka Jr.: Apparent weight of photons, Phys. Rev. Lett. **4**, 337–341 (1960)

22.72 M.M. Nieto, J.T. Goldman: The arguments against 'antigravity' and the gravitational acceleration of antimatter, Phys. Rep. **205**, 221–281 (1991)

22.73 M. Amoretti, C. Amsler, G. Bonomi, A. Bouchta, P. Bowe, C. Carraro, C.L. Cesar, M. Charlton, M.J.T. Collier, M. Doser, V. Filippini, K.S. Fine, A. Fontana, M.C. Fujiwara, R. Funakoshi, P. Genova, J.S. Hangst, R.S. Hayano, M.H. Holzscheiter, L.V. Jørgensen, V. Lagomarsino, R. Landua, D. Lindelöf, E.L. Rizzini, M. Macrì, N. Madsen, G. Manuzio, M. Marchesotti, P. Montagna, H. Pruys, C. Regenfus, P. Riedler, J. Rochet, A. Rotondi, G. Rouleau, G. Testera, A. Variola, T.L. Watson, D.P. van der Werf: Production and detection of cold anti-hydrogen atoms, Nature **419**, 456–459 (2002)

22.74 G. Gabrielse, N.S. Bowden, P. Oxley, A. Speck, C.H. Storry, J.N. Tan, M. Wessels, D. Grzonka, W. Oelert, G. Schepers, T. Sefzick, J. Walz, H. Pittner, T.W. Hänsch, E.A. Hessels: Background-free observation of cold antihydrogen with field-ionization analysis of its states, Phys. Rev. Lett. **89**, 213401 (2002)

22.75 A. Apostolakis, A. Apostolakis, E. Aslanides, G. Backenstoss, P. Bargassa, O. Behnke, A. Benelli, V. Bertin, F. Blanc, P. Bloch, P. Carlson, M. Carroll, E. Cawley, G. Chardin, M.B. Chertok, M. Danielsson, M. Dejardin, J. Derre, A. Ealet, C. Eleftheriadis, L. Faravel, W. Fetscher, M. Fidecaro, A. Filipcic, D. Francis, J. Fry, E. Gabathuler, R. Gamet, H.-J. Gerber, A. Go, A. Haselden, P.J. Hayman, F. Henry-Couannier, R.W. Hollander, K. Jon-And, P.-R. Kettle, P. Kokkas, R. Kreuger, R. Le Gac, F. Leimgruber, I. Mandic, N. Manthos, G. Marel, M. Mikuz, J. Miller, F. Montanet, A. Muller, T. Nakada, B. Pagels, I. Papadopoulos, P. Pavlopoulos, G. Polivka, R. Rickenbach, B.L. Roberts, T. Ruf, L. Sakeliou, M. Schäfer, L.A. Schaller, T. Schietinger, A. Schopper, L. Tauscher, C. Thibault, F. Touchard, C. Touramanis, C.W.E. van Eijk, S. Vlachos, P. Weber, O. Wigger, M. Wolter, D. Zavrtanik, D. Zimmerman, J.R. Ellis, N.E. Mavromatos, D.V. Nanopoulos: Tests of the equivalence principle with neutral kaons, Phys. Lett. B **452**, 425–433 (1999)

22.76 O. Bertolami, C.G. Boehmer, T. Harko, F.S.N. Lobo: Extra force in $f(R)$ modified theories of gravity, Phys. Rev. D **75**, 104016 (2007)

22.77 O. Bertolami, F.S.N. Lobo, J. Páramos: Non-minimal curvature-matter couplings in modified gravity, gr-qc/0110118 (2001)

22.78 T. Damour: Testing the equivalence principle: Why and how?, Class. Quantum Gravity **13**, A33–A42 (1996)

22.79 T. Damour: Missions spatiales en physique fondamentale, Questioning the equivalence principle, ed. by C. Borde, P. Touboul (2001), gr-qc/0109063

22.80 T. Damour, A.M. Polyakov: String theory and gravity, Gen. Relativ. Gravit. **26**, 1171–1176 (1994)

22.81 T. Damour, A.M. Polyakov: The String dilaton and a least coupling principle, Nucl. Phys. B **423**, 532–558 (1994)

22.82 R. Colella, A.W. Overhauser, S.A. Werner: Observation of gravitationally induced quantum interference, Phys. Rev. Lett. **34**, 1472–1474 (1975)

22.83 M. Kasevich, S. Chu: Measurement of the gravitational acceleration of an atom with a light-pulse atom interferometer, Appl. Phys. B **54**(5), 321 (1992)

22.84 P.G. Roll, R. Krotkov, R.H. Dicke: The Equivalence of inertial and passive gravitational mass, Annals Phys. **26**, 442–517 (1964)

22.85 V.B. Braginsky, V.I. Panov: Verification of the equivalence of inertial and gravitational mass, JETP **34**, 463 (1972)

22.86 Y. Su, B.R. Heckel, E.G. Adelberger, J.H. Gundlach, M. Harris, G.L. Smith, H.E. Swanson: New tests of the universality of free fall, Phys. Rev. D **50**, 3614–3636 (1994)

22.87 E.G. Adelberger: New tests of Einstein's equivalence principle and Newton's inverse-square law, Class. Quantum Gravity **18**(13), 2397 (2001)

22.88 J.H. Gundlach: Laboratory tests of gravity, New J. Phys. **7**, 205 (2005)

22.89 V.V. Nesvizhevsky, H. Borner, A.K. Petukhov, H. Abele, S. Baessler, F.J. Ruess, T. Stoferle, A. Westphal, A.M. Gagarsky, G.A. Petrov, A.V. Strelkov: Quantum states of neutrons in the Earth's gravitational field, Nature **415**, 297–299 (2002)

22.90 V.V. Nesvizhevsky, H. Borner, A.M. Gagarsky, G.A. Petrov, A.K. Petukhov, H. Abele, S. Baessler, T. Stoferle, S.M. Solovev: Search for quantum states of the neutron in a gravitational field: Gravitational levels, Nucl. Instrum. Meth. A **440**, 754–759 (2000)

22.91 V.I. Luschikov, A.I. Frank: Quantum effects occurring when ultracold neutrons are stored on a plane, JETP **28**, 559 (1978)

22.92 O. Bertolami, J.G. Rosa, C.M.L. de Aragão, P. Castorina, D. Zappalà: Noncommutative gravitational quantum well, Phys. Rev. D **72**, 025010 (2005)

22.93 O. Bertolami, J.G. Rosa: Quantum and classical divide: The gravitational case, Phys. Lett. B **633**, 111–115 (2006)

22.94 S. Baessler, B.R. Heckel, E.G. Adelberger, U. Schmidt, J.H. Gundlach, H.E. Swanson: Improved test of the equivalence principle for gravitational self-energy, Phys. Rev. Lett. **83**, 3585–3588 (1999)

22.95 K. Nordtvedt: Equivalence principle for massive bodies. 2. Theory, Phys. Rev. **169**, 1017–1025 (1968)

22.96 K. Nordtvedt: Lunar laser ranging reexamined: The non-null relativistic contribution, Phys. Rev. D **43**, 3131–3135 (1991)

22.97 R.K. Ulrich: The influence of partial ionization and scattering states on the solar interior structure, Astrophys. J. **258**, 404–413 (1982)
22.98 K. Nordtvedt: Solar system Eötvös experiments, Icarus **12**, 91–100 (1970)
22.99 R.B. Orellana, H. Vucetich: The principle of equivalence and the Trojan asteroids, Astron. Astroph. **200**, 248–254 (1988)
22.100 R.B. Orellana, H. Vucetich: The Nordtvedt Effect in the Trojan Asteroids, Astron. Astroph. **273**, 313–317 (1993)
22.101 T. Damour, G. Schäfer: New tests of the strong equivalence principle using binary-pulsar data, Phys. Rev. Lett. **66**, 2549–2552 (1991)
22.102 N. Wex: Pulsar timing – strong gravity clock experiments. In: *Gyros, Clocks, Interferometers ...: Testing Relativistic Gravity in Space*, Lecture Notes in Physics, Vol. 562, ed. by C. Lammerzahl, C.W.F. Everitt, F.W. Hehl (Springer, Berlin 2001) pp. 381–399
22.103 D.R. Lorimer, P.C. Carvalho Freire: New limits on the strong equivalence principle from two long – period circular – orbit binary pulsars, ASP Conf. Ser. (2004)
22.104 J.D. Anderson, J.G. Williams: Long-range tests of the equivalence principle, Class. Quantum Gravity **18**, 2447–2456 (2001)
22.105 V.A. Kostelecký, S. Samuel: Spontaneous breaking of Lorentz symmetry in string theory, Phys. Rev. D **39**, 683 (1989)
22.106 V.A. Kostelecký, S. Samuel: Phenomenological gravitational constraints on strings and higher dimensional theories, Phys. Rev. Lett. **63**, 224 (1989)
22.107 O. Bertolami: Lorentz invariance and the cosmological constant, Class. Quantum Gravity **14**, 2785–2791 (1997)
22.108 P.R. Phillips: Is the graviton a Goldstone boson?, Phys. Rev. **146**, 966–973 (1966)
22.109 V.A. Kostelecký: Gravity, Lorentz violation, and the standard model, Phys. Rev. D **69**, 105009 (2004)
22.110 V.A. Kostelecký, R.J. Van Kooten: Bounding CPT violation in the neutral B system, Phys. Rev. D **54**, 5585–5597 (1996)
22.111 R. Bluhm, V.A. Kostelecký: Spontaneous Lorentz violation, Nambu–Goldstone modes, and gravity, Phys. Rev. D **71**, 065008 (2005)
22.112 H.B. Nielsen, M. Ninomiya: Beta function in a noncovariant Yang–Mills theory, Nucl. Phys. B **141**, 153 (1978)
22.113 O. Bertolami, D.F. Mota: Primordial magnetic fields via spontaneous breaking of Lorentz invariance, Phys. Lett. B **455**, 96–103 (1999)
22.114 G. Amelino-Camelia, J.R. Ellis, N.E. Mavromatos, D.V. Nanopoulos, S. Sarkar: Tests of quantum gravity from observations of gamma-ray bursts, Nature **393**, 763–765 (1998)
22.115 S.D. Biller, A.C. Breslin, J. Buckley, M. Carson, M. Catanese, D.A. Carter-Lewis, M.F. Cawley, D.J. Fegan, J.P. Finley, J.A. Gaidos, A.M. Hillas, F. Krennrich, R.C. Lamb, R. Lessard, C. Masterson, J.E. McEnery, B. McKernan, P. Moriarty, J. Quinn, H.J. Rose, F. Samuelson, G. Sembroski, P. Skelton, T.C. Weekes: Limits to quantum gravity effects on energy dependence of the speed of light from observations of TeV flares in active galaxies, Phys. Rev. Lett. **83**, 2108–2111 (1999)
22.116 S.M. Carroll, J.A. Harvey, V.A. Kostelecký, C.D. Lane, T. Okamoto: Noncommutative field theory and Lorentz violation, Phys. Rev. Lett. **87**, 141601 (2001)
22.117 O. Bertolami, L. Guisado: Noncommutative scalar field coupled to gravity, Phys. Rev. D **67**, 025001 (2003)
22.118 O. Bertolami, L. Guisado: Noncommutative field theory and violation of translation invariance, J. High Energy Phys. **0312**, 013 (2003)
22.119 S. Imai, N. Sasakura: Scalar field theories in a Lorentz invariant three-dimensional noncommutative space-time, J. High Energy Phys. **0009**, 032 (2000)
22.120 J.M. Conroy, H.J. Kwee, V. Nazaryan: Phenomenology of Lorentz conserving noncommutative QED, Phys. Rev. D **68**, 054004 (2003)
22.121 D. Robbins, S. Sethi: The UV/IR interplay in theories with space-time varying noncommutativity, J. High Energy Phys. **0307**, 034 (2003)
22.122 R. Gambini, J. Pullin: Nonstandard optics from quantum space-time, Phys. Rev. D **59**, 124021 (1999)
22.123 J. Alfaro, H.A. Morales-Tecotl, L.F. Urrutia: Quantum gravity corrections to neutrino propagation, Phys. Rev. Lett. **84**, 2318–2321 (2000)
22.124 L.J. Garay: Space-time foam as a quantum thermal bath, Phys. Rev. Lett. **80**, 2508–2511 (1998)
22.125 J.R. Ellis, N.E. Mavromatos, D.V. Nanopoulos: Search for quantum gravity, Gen. Relativ. Gravit. **31**, 1257–1262 (1999)
22.126 V.A. Kostelecký, R. Lehnert, M.J. Perry: Spacetime – varying couplings and Lorentz violation, Phys. Rev. D **68**, 123511 (2003)
22.127 O. Bertolami, R. Lehnert, R. Potting, A. Ribeiro: Cosmological acceleration, varying couplings, and Lorentz breaking, Phys. Rev. D **69**, 083513 (2004)
22.128 D. Mattingly: Modern tests of Lorentz invariance, Living Rev. Relativ. **8**, 5 (2005)
22.129 V.A. Kostelecký, R. Potting: CPT, strings, and meson factories, Phys. Rev. D **51**, 3923–3935 (1995)
22.130 V.A. Kostelecký, R. Potting: Expectation values, Lorentz invariance, and CPT in the open bosonic string, Phys. Lett. B **381**, 89–96 (1996)
22.131 D. Colladay, V.A. Kostelecký: Tests of direct and indirect CPT violation at a B factory, Phys. Lett. B **344**, 259–265 (1995)
22.132 D. Colladay, V.A. Kostelecký: Testing CPT with the neutral D system, Phys. Rev. D **52**, 6224–6230 (1995)

22.133 R. Bluhm, V.A. Kostelecký, N. Russell: Testing CPT with anomalous magnetic moments, Phys. Rev. Lett. **79**, 1432–1435 (1997)
22.134 R. Bluhm, V.A. Kostelecký, N. Russell: CPT and Lorentz tests in Penning traps, Phys. Rev. D **57**, 3932–3943 (1998)
22.135 R. Bluhm, V.A. Kostelecký, N. Russell: CPT and Lorentz tests in hydrogen and anti-hydrogen, Phys. Rev. Lett. **82**, 2254–2257 (1999)
22.136 R. Adler, A. Angelopoulos, A. Apostolakis, E. Aslanides, G. Backenstoss, C.P. Bee, O. Behnke, A. Benelli, V. Bertin, F. Blanc, P. Bloch, P. Carlson, M. Carroll, J. Carvalho, E. Cawley, S. Charalambous, G. Chardin, M.B. Chertok, A. Cody, M. Danielsson, M. Dejardin, J. Derre, A. Ealet, B. Eckart, C. Eleftheriadis, I. Evangelou, L. Faravel, P. Fassnacht, C. Felder, R. Ferreira-Marques, W. Fetscher, M. Fidecaro, A. Filipčič, D. Francis, J. Fry, E. Gabathuler, R. Gamet, D. Garreta, H.-J. Gerber, A. Go, C. Guyot, A. Haselden, P.J. Hayman, F. Henry-Couannier, R.W. Hollander, E. Hubert, K. Jon-And, P.-R. Kettle, C. Kochowski, P. Kokkas, R. Kreuger, R. Le Gac, F. Leimgruber, A. Liolios, E. Machado, I. Mandić, N. Manthos, G. Marel, M. Mikuž, J. Miller, F. Montanet, T. Nakada, B. Pagels, I. Papadopoulos, P. Pavlopoulos, J. Pinto da Cunha, A. Policarpo, G. Polivka, R. Rickenbach, B.L. Roberts, T. Ruf, L. Sakeliou, P. Sanders, C. Santoni, M. Schäfer, L.A. Schaller, T. Schietinger, A. Schopper, P. Schune, A. Soares, L. Tauscher, C. Thibault, F. Touchard, C. Touramanis, F. Triantis, E. van Beveren, C.W.E. van Eijk, G. Varner, S. Vlachos, P. Weber, O. Wigger, M. Wolter, C. Yeche, D. Zavrtanik, D. Zimmerman, J. Ellis, J.L. Lopez, N.E. Mavromatos, D.V. Nanopoulos: Tests of CPT symmetry and quantum mechanics with experimental data from CPLEAR, Phys. Lett. B **364**, 239–245 (1995)
22.137 O. Bertolami, D. Colladay, V.A. Kostelecký, R. Potting: CPT violation and baryogenesis, Phys. Lett. B **395**, 178–183 (1997)
22.138 A. Brillet, J.L. Hall: Improved laser test of the isotropy of space, Phys. Rev. Lett. **42**, 549–552 (1979)
22.139 V.W. Hughes, H.G. Robinson, V. Beltran-Lopez: Upper limit for the anisotropy of inertial mass from nuclear resonance experiments, Phys. Rev. Lett. **4**, 342–344 (1960)
22.140 R.W.P. Drever: A search for anisotropy of inertial mass using a free precession technique, Philos. Mag. **6**, 683–687 (1961)
22.141 S.K. Lamoreaux, J.P. Jacobs, B.R. Heckel, F.J. Raab, E.N. Fortson: New limits on spatial anisotropy from optically pumped He-201 and Hg-199, Phys. Rev. Lett. **57**, 3125–3128 (1986)
22.142 V.A. Kostelecký, C.D. Lane: Constraints on Lorentz violation from clock comparison experiments, Phys. Rev. D **60**, 116010 (1999)
22.143 J.F. Bell: A tighter constraint on post-Newtonian gravity using millisecond pulsars, Astrophys. J. **462**, 287 (1996)
22.144 J.F. Bell, T. Damour: A new test of conservation laws and Lorentz invariance in relativistic gravity, Class. Quantum Gravity **13**, 3121–3128 (1996)
22.145 K. Greisen: End to the cosmic-ray spectrum?, Phys. Rev. Lett. **16**, 748–750 (1966)
22.146 G.T. Zatsepin, V.A. Kuzmin: Upper limit of the spectrum of cosmic rays, JETP Lett. **4**, 78–80 (1966)
22.147 N. Hayashida, K. Honda, M. Honda, N. Inoue, S. Imaizumi, K. Kadota, F. Kakimoto, K. Kamata, S. Kawaguchi, N. Kawasumi, Y. Matsubara, K. Murakami, M. Nagano, H. Ohoka, M. Takeda, M. Teshima, I.I. Tsushima, S. Yoshida, H. Yoshii: Observation of a very energetic cosmic ray well beyond the predicted 2.7-K cutoff in the primary energy spectrum, Phys. Rev. Lett. **73**, 3491–3494 (1994)
22.148 M. Takeda, N. Hayashida, K. Honda, N. Inoue, K. Kadota, F. Kakimoto, K. Kamata, S. Kawaguchi, Y. Kawasaki, N. Kawasumi, H. Kitamura, E. Kusano, Y. Matsubara, K. Murakami, M. Nagano, D. Nishikawa, H. Ohoka, N. Sakaki, M. Sasaki, K. Shinozaki, N. Souma, M. Teshima, R. Torii, I. Tsushima, Y. Uchihori, T. Yamamoto, S. Yoshida, H. Yoshii: Extension of the cosmic ray energy spectrum beyond the predicted Greisen–Zatsepin–Kuz'min cutoff, Phys. Rev. Lett. **81**, 1163–1166 (1998)
22.149 D.J. Bird, S.C. Corbató, H.Y. Dai, B.R. Dawson, J.W. Elbert, T.K. Gaisser, K.D. Green, M.A. Huang, D.B. Kieda, S. Ko, C.G. Larsen, E.C. Loh, M. Luo, M.H. Salamon, D. Smith, P. Sokolsky, P. Sommers, T. Stanev, J.K.K. Tang, S.B. Thomas, S. Tilav: Evidence for correlated changes in the spectrum and composition of cosmic rays at extremely high-energies, Phys. Rev. Lett. **71**, 3401–3404 (1993)
22.150 D.J. Bird, S.C. Corbato, H.Y. Dai, B.R. Dawson, J.W. Elbert, B.L. Emerson, K.D. Green, M.A. Huang, D.B. Kieda, M. Luo, S. Ko, C.G. Larsen, E.C. Loh, M.H. Salamon, J.D. Smith, P. Sokolsky, P. Sommers, J.K.K. Tang, S.B. Thomas: The cosmic ray energy spectrum observed by the Fly's Eye, Astrophys. J. **424**, 491–502 (1994)
22.151 D.J. Bird, S.C. Corbato, H.Y. Dai, J.W. Elbert, K.D. Green, M.A. Huang, D.B. Kieda, S. Ko, C.G. Larsen, E.C. Loh, M.Z. Luo, M.H. Salamon, J.D. Smith, P. Sokolsky, P. Sommers, J.K.K. Tang, S.B. Thomas: Detection of a cosmic ray with measured energy well beyond the expected spectral cutoff due to cosmic microwave radiation, Astrophys. J. **441**, 144–150 (1995)
22.152 M.A. Lawrence, R.J.O. Reid, A.A. Watson: The Cosmic ray energy spectrum above 4×10^{17} eV as measured by the Haverah Park array, J. Phys. **G17**, 733–757 (1991)

22.153 N.N. Efimov, T.A. Egorov, A.V. Glushkov, M.I. Pravdin, I.E. Sleptsov: The energy spectrum and anisotropy of primary cosmic rays at energy $E_0 > 10^{-17}$ observed in Yakutsk, ICRR Symp. Astrophys. Asp. Most Energ. Cosm. Rays, ed. by N. Nagano, F. Takahara (World Scientific, Singapore 1991)

22.154 R.U. Abbasi, T. Abu-Zayyad, J.F. Amann, G. Archbold, R.W. Atkins, J.A. Bellido, K. Belov, J.W. Belz, S. BenZvi, D.R. Bergman, G.W. Burt, Z. Cao, R.W. Clay, B. Connolly, W. Deng, B.R. Dawson, Y. Fedorova, J. Findlay, C.B. Finley, W.F. Hanlon, C.M. Hoffman, G.A. Hughes, M.H. Holzscheiter, P. Huntemeyer, C.C.H. Jui, K. Kim, M.A. Kirn, E.C. Loh, M.M. Maestas, N. Manago, L.J. Marek, K. Martens, J.A.J. Matthews, J.N. Matthews, A. O'Neill, C.A. Painter, L. Perera, K. Reil, R. Riehle, M. Roberts, M. Sasaki, S.R. Schnetzer, K.M. Simpson, G. Sinnis, J.D. Smith, R. Snow, P. Sokolsky, C. Song, R.W. Springer, B.T. Stokes, J.R. Thomas, S.B. Thomas, G.B. Thomson, D. Tupa, S. Westerhoff, L.R. Wiencke, A. Zech: Observation of the ankle and evidence for a high-energy break in the cosmic ray spectrum, Phys. Lett. B **619**, 271–280 (2005)

22.155 Pierre Auger Observatory: http://www.auger.org/icrc2005/

22.156 H. Sato, T. Tati: Hot universe, cosmic rays of ultrahigh energy and absolute reference system, Prog. Theor. Phys. **47**, 1788–1790 (1972)

22.157 S.R. Coleman, S.L. Glashow: Cosmic ray and neutrino tests of special relativity, Phys. Lett. B **405**, 249–252 (1997)

22.158 S.R. Coleman, S.L. Glashow: High-energy tests of Lorentz invariance, Phys. Rev. D **59**, 116008 (1999)

22.159 L. Gonzalez-Mestres: Deformed Lorentz symmetry and ultrahigh-energy cosmic rays, 26th Int. Cosm. Ray Conf. (ICRC 1999) (1999), hep-ph/9905430

22.160 O. Bertolami: Ultrahigh-energy cosmic rays and symmetries of space-time, Gen. Relativ. Gravit. **34**, 707–713 (2002)

22.161 O. Bertolami: Threshold effects and Lorentz symmetry, Lecture Notes in Physics **633**, 96–102 (2003)

22.162 R.F.C. Vessot, M.W. Levine, E.M. Mattison, E.L. Blomberg, T.E. Hoffman, G.U. Nystrom, B.F. Farrel, R. Decher, P.B. Eby, C.R. Baugher, J.W. Watts, D.L. Teuber, F.D. Wills: Test of relativistic gravitation with a space-borne hydrogen maser, Phys. Rev. Lett. **45**, 2081–2084 (1980)

22.163 N. Ashby, T.P. Heavner, S.R. Jefferts, T.E. Parker, A.G. Radnaev, Y.O. Dudin: Testing local position invariance with four cesium-fountain primary frequency standards and four NIST hydrogen masers, Phys. Rev. Lett. **98**, 070802 (2007)

22.164 J.D. Anderson, P.A. Laing, E.L. Lau, A.S. Liu, M.M. Nieto, S.G. Turyshev: Indication, from Pioneer 10/11, Galileo, and Ulysses data, of an apparent anomalous, weak, long range acceleration, Phys. Rev. Lett. **81**, 2858–2861 (1998)

22.165 J.D. Anderson, P.A. Laing, E.L. Lau, A.S. Liu, M.M. Nieto, S.G. Turyshev: Study of the anomalous acceleration of Pioneer 10 and 11, Phys. Rev. D **65**, 082004 (2002)

22.166 C.B. Markwardt: Independent confirmation of the Pioneer 10 anomalous acceleration, gr-qc/0208046 (2002)

22.167 V.T. Toth: Independent analysis of the orbits of Pioneer 10 and 11, Int. J. Mod. Phys. D **18**, 717–741 (2009)

22.168 A. Levy, B. Christophe, P. Berio, J.-M. Courty, G. Metris, S. Reynaud: Pioneer Doppler data analysis: Study of periodic anomalies, Adv. Space Res. **43**, 1538–1544 (2009)

22.169 J.I. Katz: Comment on 'Indication, from Pioneer 10/11, Galileo, and Ulysses data, of an apparent anomalous, weak, long range acceleration', Phys. Rev. Lett. **83**, 1892 (1999)

22.170 L.K. Scheffer: Conventional forces can explain the anomalous acceleration of Pioneer 10, Phys. Rev. D **67**, 084021 (2003)

22.171 O. Bertolami, J. Páramos: The Pioneer anomaly in a bimetric theory of gravity on the brane, Class. Quantum Gravity **21**, 3309–3321 (2004)

22.172 M.-T. Jaekel, S. Reynaud: Post-Einsteinian tests of linearized gravitation, Class. Quantum Gravity **22**, 2135–2158 (2005)

22.173 J.R. Brownstein, J.W. Moffat: Gravitational solution to the Pioneer 10/11 anomaly, Class. Quantum Gravity **23**, 3427–3436 (2006)

22.174 O. Bertolami, P. Vieira: Pioneer anomaly and the Kuiper Belt mass distribution, Class. Quantum Gravity **23**, 4625–4635 (2006)

22.175 M.M. Nieto: Analytic gravitational-force calculations for models of the Kuiper belt, with application to the Pioneer anomaly, Phys. Rev. D **72**, 083004 (2005)

22.176 O. Bertolami, F. Francisco, P.J.S. Gil, J. Páramos: Thermal analysis of the Pioneer anomaly: A method to estimate radiative momentum transfer, Phys. Rev. D **78**, 103001 (2008)

22.177 S.G. Turyshev, V.T. Toth: The Pioneer anomaly in the light of new data, Space Sci. Rev. **148**, 149 (2009)

22.178 B. Rievers, S. Bremer, M. List, C. Lämmerzahl, H. Dittus: Thermal dissipation force modeling with preliminary results for Pioneer 10/11, Acta Astronaut. **66**(3-4), 467 (2010)

22.179 B. Rievers, C. Lämmerzahl, M. List, S. Bremer, H. Dittus: New powerful thermal modelling for high-precision gravity missions with application to Pioneer 10/11, New J. Phys. **11**, 113032 (2009)

22.180 O. Bertolami, F. Francisco, P.S.J. Gil, J. Páramos: Estimating radiative momentum transfer through a thermal analysis of the Pioneer anomaly, Space Sci. Rev. **151**, 75–91 (2010)

22.181 F. Francisco, O. Bertolami, P.J.S. Gil, J. Páramos: Modelling the reflective thermal contribution to

the acceleration of the Pioneer spacecraft, Phys. Lett. B **711**, 337–346 (2012)

22.182 B. Rievers, C. Lammerzahl: High precision thermal modeling of complex systems with application to the flyby and Pioneer anomaly, Ann. Phys. **523**, 439–449 (2011)

22.183 S.G. Turyshev, V.T. Toth, G. Kinsella, S.-C. Lee, S.M. Lok, J. Ellis: Support for the thermal origin of the Pioneer anomaly, Phys. Rev. Lett. **108**, 241101 (2012)

22.184 J.D. Anderson, J.K. Campbell, J.E. Ekelund, J. Ellis, J.F. Jordan: Anomalous Orbital-Energy Changes Observed during Spacecraft Flybys of Earth, Phys. Rev. Lett. **100**, 091102 (2008)

22.185 P.G. Antreasian, J.R. Guinn: Investigations into the unexpected delta-V increases during the earth gravity assists of Galileo and NEAR, AIAA/AAS Astrodyn. Spec. Conf. Exhib. (1998), No. 98-4287

22.186 S.L. Adler: Modeling the flyby anomalies with dark matter scattering: update with additional data and further predictions, astro-ph.EP/1112.5426 (2011)

22.187 C. Laemmerzahl, O. Preuss, H. Dittus: Is the physics within the Solar system really understood? In: *Lasers, Clocks, and Drag-Free: Technologies for Future Exploration in Space and Tests of Gravity*, ed. by C. Laemmerzahl, O. Preuss, H. Dittus (Springer, Berlin 2006), gr-qc/0604052

22.188 N. Ashby: Relativity in the global positioning system, Living Rev. Relativ. **6**, 1 (2003)

22.189 H. Busack: Test for consistence of a flyby anomaly simulation with the observed Doppler residuals for the Messenger flybys of Mercury (2010), physics.gen-ph/1006.3555

22.190 O. Bertolami, F. Francisco, P.J.S. Gil, J. Páramos: Testing the flyby anomaly with the GNSS constellation, Int. J. Mod. Phys. D **21**, 1250035 (2012)

22.191 J. Páramos, G. Hechenblaikner: Probing the flyby anomaly with the future STE-QUEST mission, gr-qc/1210.7333 (2012)

22.192 O. Bertolami: The Cosmological constant problem: A user's guide, Int. J. Mod. Phys. D **18**, 2303–2310 (2009)

22.193 E.P. Verlinde: On the origin of gravity and the laws of Newton, J. High Energy Phys. **1104**, 029 (2011)

22.194 C. Bastos, O. Bertolami, N.C. Dias, J.N. Prata: Entropic gravity, phase-space noncommutativity and the equivalence principle, Class. Quantum Gravity **28**, 125007 (2011)

23. Observational Constraints on Local Lorentz Invariance

Robert T. Bluhm

Local Lorentz invariance is a fundamental spacetime symmetry in the standard model of particle physics and in general relativity. However, in a quantum theory of gravity, mechanisms have been found to arise that might allow small violations of Lorentz invariance to occur. An effective field theory known as the standard model extension has been developed to search for these violations. The standard model extension incorporates Lorentz-violating interaction terms involving particle fields and gravitational fields, and it includes all terms that could arise from a process of spontaneous Lorentz violation as well as terms that explicitly break Lorentz symmetry. In this chapter, an overview of the standard model extension is presented, including its motivations and construction. A partial survey of high-precision experimental tests of local Lorentz invariance for the different particle sectors in the standard model and with gravity is presented as well.

The standard model (SM) of particle physics and the theory of general relativity (GR) are currently the best theories describing the four fundamental forces of nature: electromagnetism, strong and weak nuclear forces, and gravity. There are no known experimental conflicts with predictions from either of these theories. Nonetheless, they are fundamentally different in that the SM is a quantum theory, while GR is a classical geometrical theory. It remains an open issue as to how to merge or reconcile the SM and GR into a unified theory that presumably contains a quantum description for gravity. The relevant scale for a quantum theory of gravity is typically taken as the Planck scale, which is approximately 10^{19} GeV. Promising candidates for a quantum theory of gravity include string theory and loop quantum gravity. These and other ideas for quantizing gravity can involve new features such as, for example, higher dimensions of space and time, braneworld scenarios, noncommutative geometries, and spacetime-varying fields or couplings. It is also possible that in merging GR into a quantum theory of gravity, the laws of relativity might not hold exactly at all energy scales.

Searching for experimental evidence of a quantum theory of gravity is challenging because conducting experiments at the Planck scale is not possible. However, suppressed effects emerging from a more fundamental theory might be observable in highly-sensitive low-energy experiments or in interferometry experiments with extremely long baselines. One candidate set of Planck-scale-suppressed signals is relativity violations associated with small breaking of local Lorentz symmetry. It has been shown, for example, that mechanisms arising in the context of string theory and other quantum theories of gravity might lead to violation of Lorentz symmetry. It is for this reason that considerable inter-

est in the possibility of Lorentz violation has emerged in recent years, and a number of new high-precision experimental tests of local Lorentz invariance have been performed.

A key development in the investigation of Lorentz violation was the formulation of a comprehensive theoretical framework known as the standard model extension (SME) [23.1–5]. It contains both the SM and generalized theories of gravity (including GR) as well as all possible observer-independent Lorentz-violating interactions involving particle and gravitational fields. The SME has been used extensively to test for Lorentz violation in high-precision experiments. It also has theoretical features that are important for understanding the different types of processes that might lead to Lorentz violation.

In this overview, the focus is on using the SME to investigate the possibility of Lorentz violation both theoretically and experimentally. An underlying assumption in using the SME is that at low energies (compared to the Planck scale) Lagrangian-based field theory gives what is currently the best description of elementary particles and their interactions. Therefore, if some new type of physics, such as Lorentz violation, goes beyond the SM and GR, then according to this assumption its leading-order corrections should be describable in the context of effective field theory. It is for this reason that the SME is suitable as a framework for investigating signals of Lorentz violation in experiments. However, in the search for a consistent quantum theory of gravity, it is also possible to consider ideas that fall outside the domain of Lagrangian-based field theory. These include ideas such as the breakdown of quantum mechanics, or where spacetime becomes discrete or noncontinuous at the quantum-gravity scale. Many of these ideas also lead to Lorentz violation. To the extent that they can be described at the level of effective field theory at low energies, then they should give rise to effects that are contained in the SME framework. However, if these alternative theories cannot be described using Lagrangian-based effective field theory, then it may not be possible to investigate their effects using the SME. In those cases, one would need to work within the context of the given model in order to investigate possible signals of Lorentz violation.

23.1 Spacetime Symmetries in Relativity

In special relativity, the equations of motion for particles and fields are invariant under Lorentz transformations. The Lorentz symmetry in this case is a global symmetry, with the transformations being the same at each point in the spacetime. The geometry of special relativity is that of a flat spacetime or Minkowski spacetime. In contrast, in GR, the effects of gravity are described by the curvature of spacetime, and the geometry is Riemannian. Lorentz symmetry still holds in GR, but only locally, e.g., in instantaneous infinitesimal inertial frames. In these local frames, called local Lorentz frames, the laws of special relativity are assumed to hold according to the equivalence principle. The symmetry in this case is local Lorentz invariance (LLI).

Curved spacetime in GR is described by the metric tensor, $g_{\mu\nu}$, the Riemann curvature tensor, $R^{\kappa}{}_{\lambda\mu\nu}$, and its contractions, including the Ricci tensor, $R_{\mu\nu}$, and the curvature scalar R. These quantities appear in the Einstein field equations along with the energy-momentum tensor for the matter fields, $T_{\mathrm{M}}^{\mu\nu}$, which acts as the source of the spacetime curvature. The Einstein equations are invariant under a set of spacetime transformations (defined mathematically in a later section) known as diffeomorphisms, which consist of mappings of the curved spacetime manifold back onto itself.

In particle physics described using special relativity, the matter fields often have additional symmetries, such as internal gauge symmetry or discrete spacetime symmetries. The latter include parity, P, charge conjugation, C, and time reversal, T. In the SM of particle physics, many of these symmetries are broken either explicitly or through a process of spontaneous symmetry breaking. For example, spontaneous breaking of gauge symmetry is an essential feature of the Higgs mechanism in the electroweak model. In addition, all of the discrete symmetries C, P, and T are broken by the weak interactions, including the combination CP, which is broken in certain meson interactions. However, a theorem in particle physics, known as the CPT theorem, states that the combination of all three of the discrete spacetime symmetries, CPT, must hold for all local interactions of point-like particles in the context of quantum field theory [23.6–9]. An essential assumption of the CPT theorem is that Lorentz symmetry must hold. This is important in investigations of Lorentz violation because it implies that if Lorentz symmetry is

broken, then CPT breaking could occur as well because the conditions for the theorem to hold would not apply. This opens up another avenue of investigation of Lorentz violation in that high-precision tests of CPT symmetry can be used as well to test for Lorentz violation. Another theorem in the context of quantum field theory strengthens this connection. It states that in realistic effective field theories, interactions that break CPT also break Lorentz symmetry [23.10]. Evidently, there is a strong link between CPT violation and Lorentz violation, and any experiment looking for CPT violation can also be viewed as a Lorentz test in the context of quantum field theory.

The SM of particle physics can be combined with GR to describe all four of the fundamental forces. This involves using a curved background with a metric $g_{\mu\nu}$ to describe the physical spacetime in which the SM particles move and interact. The resulting theory is a hybrid theory in which the SM fields are quantum fields with local $SU(3) \times SU(2) \times U(1)$ gauge symmetry. However, the metric field is not quantized, and the pure gravity sector of the theory remains a classical theory. A classical Lagrangian can be written down for the full theory as a sum of an SM sector and a gravity sector,

$$\mathcal{L} = \mathcal{L}_{\mathrm{SM}} + \mathcal{L}_{\mathrm{GR}} \,. \tag{23.1}$$

Derivatives of fields included in these expressions must be both gauge covariant and gravitationally covariant. The latter involves the introduction of a spacetime connection $\Gamma^{\lambda}{}_{\mu\nu}$. The classical action of the theory is then given by the integral

$$S = \int \sqrt{-g}\mathcal{L}\,\mathrm{d}^4x \,, \tag{23.2}$$

where the factor involving the determinant of the metric, g, ensures that the spacetime volume element in the integral is covariant under general coordinate transformations. The Einstein field equations are obtained by variation of the action with respect to the metric. The SM fields appear in the Einstein equations by contributing to the energy-momentum tensor for the matter fields, $T^{\mu\nu}_{\mathrm{M}}$. Both the action and the field equations are invariant under diffeomorphism transformations.

In most particle physics experiments, the gravitational interactions are irrelevant. In this case, the contributions from $\mathcal{L}_{\mathrm{GR}}$ are dropped and the metric is set equal to the Minkowski metric, $\eta_{\mu\nu}$. The theory can then be treated in the context of quantum field theory, where special relativity alone suffices. Without gravity, the relevant symmetries are global Lorentz symmetry and the local gauge symmetry of the SM, $SU(3) \times SU(2) \times U(1)$. However, with gravity included, the relevant symmetries of the theory change. Lorentz symmetry becomes a local symmetry, and diffeomorphism symmetry becomes important as well.

To observe the LLI of a theory in a curved spacetime, one approach is to make a coordinate transformation to a local Lorentz frame at each point in the spacetime manifold. In this way, the metric in the local coordinate system at each point reduces to $\eta_{\mu\nu}$, the connection vanishes, and locally the laws of special relativity apply. The choice of local Lorentz frame is not unique, however, since a Lorentz transformation at a given point leaves $\eta_{\mu\nu}$ unchanged.

An alternative approach keeps the spacetime frame fixed with metric $g_{\mu\nu}$, but also reveals the LLI at the same time. In this approach, four vector fields, $e_\mu{}^a$, with labels $a = 0, 1, 2, 3$ are introduced. They are called vierbein or tetrad fields. They relate tensor components in the space-time frame (labeled by Greek indices) to the corresponding components in a local Lorentz frame (labeled by Latin indices). For example, the metric obeys

$$g_{\mu\nu} = e_\mu{}^a e_\nu{}^b \eta_{ab} \,. \tag{23.3}$$

Since the metric is a symmetric field obeying $g_{\mu\nu} = g_{\nu\mu}$, it has at most ten independent degrees of freedom. In contrast, the vierbein, $e_\mu{}^a$, has a total of sixteen independent degrees of freedom. The six extra degrees of freedom are associated with the LLI.

There are several advantages to studying possible violations of LLI in a vierbein formalism. One is that fermions can more readily be introduced. In GR, particles form tensor representations under the group of linear transformations associated with general coordinate transformations, and there are no representations for spin-half fermions. Thus, it is not possible to define Dirac gamma matrices or covariant derivatives of spinor fields in a spacetime manifold in GR. However, with a vierbein formalism it is possible to extend the usual definitions of these quantities in special relativity into curved spacetime. Another advantage of a vierbein formalism is that it allows the local Lorentz symmetry and diffeomorphism symmetry to be treated in a manner similar to local gauge symmetry in particle physics. However, to do this in a general way requires that an additional geometrical quantity called torsion be introduced into the theory. Geometrically, theories with torsion allow a twisting of coordinate axes as the axes are transported along a curve. This twisting cannot be described by the curvature tensor alone. The resulting

geometry when torsion is included is called Riemann–Cartan geometry. (For reviews describing torsion and Riemann–Cartan geometry, see [23.11, 12]). For these reasons, many investigations of Lorentz violation use a vierbein formalism and work in a generalized geometry, such as Riemann–Cartan geometry.

The use of a vierbein also involves the introduction of a spin connection. It enters in covariant derivatives acting on local tensor components and plays the role of the gauge field for the Lorentz symmetry. In contrast, excitations of the metric field can be viewed as the gauge fields for the diffeomorphism symmetry. The relationship between the vierbein and spin connection is often a reflection of the type of spacetime geometry being considered. For example, in a Riemannian geometry (with no torsion), the spin connection is nondynamical and does not propagate. However, in a Riemann–Cartan geometry (with nonzero torsion), the spin connection must be treated as independent degrees of freedom that can propagate in principle. These different types of geometry can have effects on mechanisms that occur when Lorentz symmetry is violated. This is especially the case when Lorentz symmetry is spontaneously broken.

23.1.1 Lorentz Transformations and Diffeomorphisms

In special relativity, the Lorentz transformations consist of three rotations and three boosts. They are constant linear transformations that leave the Minkowski metric, $\eta_{\mu\nu}$, invariant. Mathematically, they can be implemented by contracting the tensors in a theory with a transformation matrix $\Lambda_\mu{}^\alpha$. In Cartesian coordinates, the transformation matrix obeys,

$$\eta_{\mu\nu} = \Lambda_\mu{}^\alpha \Lambda_\nu{}^\beta \eta_{\alpha\beta} \,. \tag{23.4}$$

It is often useful to consider infinitesimal Lorentz transformations, which can be written as $\Lambda_\mu{}^\alpha \simeq \delta_\mu{}^\alpha + \epsilon_\mu{}^\alpha$, where the six parameters, $\epsilon_\mu{}^\alpha = -\epsilon^\alpha{}_\mu$, generate infinitesimal rotations and boosts. Under an infinitesimal particle Lorentz transformation, a tensor $T^{\lambda\mu\nu\cdots}$ transforms as,

$$T^{\lambda\mu\nu\cdots} \to T^{\lambda\mu\nu\cdots} + \epsilon^\lambda{}_\rho T^{\rho\mu\nu\cdots} + \epsilon^\mu{}_\rho T^{\lambda\rho\nu\cdots} + \epsilon^\nu{}_\rho T^{\lambda\mu\rho\cdots} + \cdots \,. \tag{23.5}$$

In a theory with LLI, the action describing the theory and the equations of motion are left unchanged when all of the tensor fields in the theory are transformed by infinitesimal Lorentz transformations.

In the presence of gravity, the vierbein can be used to relate tensor components in a local Lorentz frame to the corresponding components in the space-time frame. A vierbein field appears for each tensor index. For example, for the tensor $T^{\lambda\mu\nu\cdots}$,

$$T^{\lambda\mu\nu\cdots} = e^\lambda{}_a e^\mu{}_b e^\nu{}_c \cdots T^{abc\cdots} + \cdots \,. \tag{23.6}$$

A local Lorentz transformation acts on the tensor components defined with respect to the local frame, e.g., $T^{abc\cdots}$. For a local infinitesimal transformation, the six Lorentz parameters are written as ϵ_{ab}. These depend on the spacetime coordinates at a given point. Under a local Lorentz transformation, the vierbein transforms as a vector,

$$e_\mu{}^a \to e_\mu{}^a + \epsilon^a{}_d e_\mu{}^d \,. \tag{23.7}$$

Typically, in a gravitational theory with LLI, the six degrees of freedom associated with the local Lorentz symmetry are used to gauge away the six antisymmetric components in the vierbein. The remaining ten components are symmetric and can be written in terms of field excitations $h_\mu{}^a = h^a{}_\mu$. For small excitations about a flat Minkowski background, the form of the vierbein can then be written as

$$e_\mu{}^a = \delta_\mu{}^a + \tfrac{1}{2} h_\mu{}^a \,. \tag{23.8}$$

Substituting this into (23.3) yields the usual expression for the metric in terms of small excitations about a Minkowski background, $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$.

Diffeomorphisms are mappings from one differentiable manifold to another. In GR, the mappings are from the spacetime manifold back to itself. Vectors and tensors transform in prescribed ways under diffeomorphisms, and diffeomorphism invariance in GR is the statement that the same physics is described by the spacetime manifold, metric, and matter fields both before and after a diffeomorphism is performed.

As with local Lorentz symmetry, diffeomorphism symmetry can be used to eliminate additional degrees of freedom. Under infinitesimal diffeomorphism transformations, points x^μ on the space-time manifold are mapped to neighboring points $x^\mu + \xi^\mu$, where the four parameters ξ^μ are spacetime dependent. Under infinitesimal diffeomorphisms, the metric transforms as

$$g_{\mu\nu} \to g_{\mu\nu} - \partial_\mu \xi_\nu - \partial_\nu \xi_\mu \,. \tag{23.9}$$

By gauge fixing the four diffeomorphism degrees of freedom, the metric can be reduced from ten down to six independent degrees of freedom. The excitations $h_{\mu\nu}$ then have six degrees of freedom as well after gauge fixing. These represent the six possible excitation modes for gravitational radiation that can occur in a generalized theory of gravity. For the case of Einstein's GR, the kinetic terms in the action are chosen so that four of these degrees of freedom do not propagate as physical modes and instead are called auxiliary modes. As a result, in Einstein's GR only two gravitational modes propagate, which are both massless transverse modes.

23.1.2 Particle and Observer Transformations

In investigations of possible Lorentz violation, it is important to distinguish between *observer* and *particle* transformations [23.2, 3]. Observer transformations are essentially changes of coordinate systems, where the tensors describing particles and fields in the system are left physically unchanged. Lorentz transformations that transform between different local or global inertial frames are examples of observer transformations. Alternatively, Lorentz transformations can be performed directly on the tensor fields in a system, while leaving the observer frame (coordinate system) unchanged. When performed this way, the transformations are called particle transformations.

Similarly, in GR, general coordinate transformations can be performed, which correspond to a change of observer frame. These are observer transformations, which leave the equations of motion covariant in form. In contrast, diffeomorphisms are particle transformations performed with respect to a fixed observer, or in a fixed coordinate frame. The particles and fields of the system, including for example the metric, are transformed under diffeomorphisms in a prescribed way that leaves the equations of motion unchanged.

It is common to hear observer and particle transformations referred to, respectively, as passive and active transformations. In theories without spacetime symmetry breaking, these transformations are essentially inverses of each other in terms of how they act on tensor quantities. However, when a symmetry is broken, this is no longer the case, and it is important to make a clearer distinction between these two types of transformations.

It is reasonable to assert that a physical interaction should not depend on the choice of coordinates of a particular observer. As a result, any physical theory should be invariant under the relevant set of observer transformations for that theory. It is for this reason that observer transformations are not particularly meaningful as symmetry transformations. The physically important symmetry transformations are the particle transformations, which can be performed in a fixed arbitrary observer frame.

Even in a theory with interactions that break a spacetime symmetry, the resulting physical description should still not depend on any particular observer or choice of coordinates. Thus, in theories with broken Lorentz symmetry, the Lagrangian and equations of motion should be unchanged when an observer Lorentz transformation is performed. However, with Lorentz-violating interactions, the particle transformations are no longer symmetries of the theory. In a given observer frame, the physics can therefore change when a particle or field is transformed under a particle Lorentz transformation.

For example, consider a scattering experiment in special relativity. If Lorentz symmetry is violated, there may be preferred spatial orientations or speeds for the incoming and outgoing particles. As a result, particles scattered in different directions or with different speeds, with respect to a given observer, may behave differently. Nonetheless, the theory remains fully observer-independent. If a different observer measures the same scattering events, the resulting physical effects will be unaffected. All that happens in an observer Lorentz transformation is that the same physical events are expressed with respect to a different Lorentz frame.

According to this approach, when LLI is broken, it is only the particle Lorentz transformations that are broken. The theory remains Lorentz observer-independent at all times. Likewise, in an extension of GR that incorporates $pacetime symmetry breaking, the relevant transformations are particle local Lorentz transformations and diffeomorphisms. If either of these are broken, the theory should still be covariant under observer local Lorentz transformations and observer general coordinate transformations.

23.1.3 Lorentz Violation

Lorentz symmetry is fundamental in both the SM and in GR. It should, therefore, be tested as accurately as possible as a way of testing the validity of these theories. In addition, it has been shown in the context of quantum-gravity theories that small violations of Lorentz symmetry might occur. For example, in string field theory mechanisms can occur that might lead to

spontaneous breaking of Lorentz symmetry [23.13–19]. Indeed, it was this idea that led to the development of the SME, which, in turn, has stimulated a variety of new experimental tests of LLI. (For reviews of various experimental and theoretical approaches to Lorentz and CPT violation see [23.20–23]).

In string field theory, a string state can be expanded as a sum of tensor-valued particle states, where the particle masses increase with the order of the tensor. String interactions provide couplings between the particle states. Spontaneous Lorentz violation can occur in this context when a string field theory has a nonperturbative vacuum that can lead to one or more of the tensor-valued fields, T, acquiring nonzero vacuum expectation values or vevs, $\langle T\rangle \neq 0$. When this occurs, the low-energy effective theory can contain terms of the form

$$\mathcal{L} \approx \frac{\lambda}{m_{\mathrm{P}}^{k}} \langle T\rangle \Gamma \bar{\psi} (\mathrm{i}\partial)^{k} \chi\,, \tag{23.10}$$

where k is an integer power, λ is a coupling constant, Γ is a generalized Dirac matrix, m_{P} is the Planck mass, and ψ and χ are fermion fields. Note that the higher-dimensional ($k > 0$) derivative couplings are expected to be balanced by additional inverse factors of the Planck mass m_{P}.

In this expression, the tensor vev, $\langle T\rangle$, carries spacetime indices, which are not written out in this notation. This vev is effectively a set of background functions or constants that are fixed in a given observer frame. As tensor-valued backgrounds, these coefficients can have preferred directions in spacetime or velocity dependence. In other words, they induce Lorentz violation. A more general interaction term can be defined by absorbing all of the couplings and inverse mass factors into the vev. The Lorentz-violating interactions then have the form

$$\mathcal{L} \approx t^{(k)} \Gamma \bar{\psi} (\mathrm{i}\partial)^{k} \chi\,, \tag{23.11}$$

where the coefficients $t^{(k)}$ carry spacetime indices and act as fixed Lorentz-violating background fields. In addition to interactions with fermions, additional terms involving gauge-field couplings and gravitational interactions are possible as well.

The SME is a generalization of these types of interactions to include all possible contractions of known SM and gravitational fields with fixed background coefficients $t^{(k)}$ [23.1–5]. This includes all arbitrary dimension interaction terms inducing Lorentz violation in effective field theory. The coefficients for Lorentz violation, $t^{(k)}$, are examples of SME coefficients. They are assumed to be heavily suppressed, presumably by inverse powers of the Planck mass. In fact, since no Lorentz violation has been observed in nature, these SME coefficients must be small.

By developing the SME in this generalized way, a framework that is particularly well suited for phenomenology results. In this approach, the Lorentz-violating SME coefficients are treated as quantities to be bounded in experiments. They can be thought of as vevs arising in a process of spontaneous Lorentz violation or simply as being due to explicit Lorentz violation from some unknown mechanism.

The interactions in (23.10) can also be used to study other processes related to Lorentz and CPT violation. For example, terms of this form have been shown to induce a form of CPT-violating baryogenesis [23.24].

Another example of Lorentz violation comes from noncommutative field theory [23.25]. These are theories with noncommuting coordinates $[x^{\mu}, x^{\nu}] = \mathrm{i}\theta^{\mu\nu}$. It has been shown that this type of geometry can occur naturally in string theory and that it leads to Lorentz violation [23.26–30]. The fixed parameters $\theta^{\mu\nu}$ break Lorentz symmetry and act effectively as fixed background tensors. For example, in an effective field theory with a $U(1)$ gauge field in a noncommutative geometry, interaction terms of the form

$$\mathcal{L} \approx \mathrm{i}q\theta^{\alpha\beta} F_{\alpha\beta} \bar{\psi} \gamma^{\mu} D_{\mu} \psi \tag{23.12}$$

can arise. Here, $F_{\alpha\beta}$ is the field strength, q is the charge, and D_{μ} is a gauge-covariant derivative. As in (23.12) the interaction takes the form of a scalar-valued product of known particle fields, derivative operators, and a set of fixed background functions. It is straightforward to write these interactions in terms of SME couplings.

There are a number of other examples of theories with Lorentz violation that have been put forward in recent years. These include models with spacetime-varying fields, quantum gravity models, multiverses, and braneworld scenarios. See, for example, [23.31–43]. It is also possible to construct models with specific types of Lorentz violation. These include models that maintain spatial rotational invariance while breaking only boost transformations, models with Lorentz-violating dispersion relations constructed using higher-order derivative interactions, and vector-tensor models in gravity that spontaneously break Lorentz symmetry. To the extent that these types of models can be described wholly or in part using Lagrangian-based ef-

fective field theory, they can be investigated using the SME. However, some ideas for quantum-gravity theories contain new features that are not readily described in the context of effective field theory. Examples include ideas such as spacetime foam, causal sets, and relative locality. See, for example, [23.44–47]. In these types of models, many of the signals of Lorentz violation that arise are not suitable for investigation using the SME and instead must be studied in the context of the specific theory.

A number of phenomenological frameworks involving certain kinds of Lorentz violation have been used by experimentalists in the past. These include the Robertson–Mansouri–Sexl framework and the PPN formalism [23.48–50]. In some cases, these and other theories describe parameterized equations of motion or dispersion relations that do not originate from a scalar Lagrangian. However, to the extent that these models can be described by effective field theory defined by a scalar Lagrangian, they are compatible with the SME and direct links between their parameterizations and the SME coefficients can be obtained. Since CPT violation in field theory is associated with Lorentz violation, it follows as well that any observer-independent effective field theory describing CPT violation should also be contained within the SME. Since CPT can be tested to very high precision in experiments comparing matter and antimatter, this type of experiment is also ripe as a testing ground for Lorentz violation.

23.2 Standard Model Extension (SME)

Currently, there is no consistent quantum theory of gravity that can be used in detailed examinations of the phenomenology of Lorentz violation at accessible energies. Nonetheless, progress can still be made using effective field theory. To be realistic, an effective field theory must contain the SM and a theory of gravity (such as GR), and it must be compatible with observations. It must also maintain observer independence. The standard model extension (SME) is defined to be the most general effective field theory of this type incorporating arbitrary observer-independent Lorentz violation.

The SME Lagrangian by definition contains all observer-scalar terms that consist of products of the SM and gravitational fields with each other as well as with additional couplings that introduce violations of Lorentz symmetry. In principle, there are an infinity of terms in the SME, including nonrenormalizable terms of arbitrary dimension. Most of these terms are expected to be suppressed by large inverse powers of the Planck scale. The question of how to extract a useful finite subset of terms from the full SME to analyze a given experiment becomes relevant, and there are a number of different ways to proceed. Perhaps the most natural approach is to follow the direction indicated by the experiments testing Lorentz violation. While these tend to be highly interdisciplinary, and include experiments in astrophysics, gravity, atomic, nuclear, and particle physics, as well as laboratory experiments with macroscopic media and space-based tests, several primary divisions and classifications can be made.

One important split is between experiments that can ignore the effects of gravity from those that cannot. For this reason, a distinction is made between limits of the SME that do not include gravity (where special relativity and global Lorentz invariance are paramount) from those where gravity is included (where Lorentz symmetry acts as a local symmetry in a curved spacetime). It is expected that the nongravitational limits of the SME will, in general, be subsets of larger SME limits that include gravity. For example, if the curvature is set to zero and the metric is replaced by the Minkowski metric, an SME limit with gravity should reduce to a corresponding SME limit in which gravity is excluded. Starting from the ground up in constructing explicit limits of the SME, it is, therefore, natural to ignore gravity at first and then to generalize the resulting theories to incorporate gravity.

In the absence of gravity, a second primary division between subsets of the SME can be made based on the types of SM fields and interactions (especially their dimensionality) that are included. Since the SM itself is a renormalizable and gauge-invariant theory, a first step in constructing a useful SME limit is to incorporate Lorentz violation while maintaining these features. This limit restricting the SME to power-counting renormalizable and gauge-invariant terms is called the minimal SME (mSME). An advantage of working with the mSME is that each particle sector has a finite independent set of mSME coefficients that can be probed experimentally. Indeed, in recent years, experimentalists have adopted using bounds on mSME coefficients

as the primary means of reporting sensitivity of their experiments to Lorentz violation.

Many of the low-energy experiments testing Lorentz violation involve only electromagnetic interactions between charged particles and photons. For this reason, it is useful as well to define a minimal QED sector of the SME. In a field theory with charged fermions, the minimal QED Lagrangian consists of the standard Dirac and Maxwell terms supplemented by Lorentz-violating terms that maintain $U(1)$ gauge symmetry and power-counting renormalizability.

If leading order effects are of primary interest, then SME limits at the level of relativistic quantum mechanics can be constructed. This is particularly useful in investigations of low-energy atomic systems, where small corrections to atomic energy levels can result from Lorentz breaking at leading order. Experiments using particle traps, masers, and high-precision spectroscopy can then be analyzed in a straightforward manner using perturbation theory.

On the other hand, if first-order effects can be ruled out in an experiment, it will be necessary to construct limits of the SME that include nonrenormalizable terms. In some scenarios for Lorentz violation, it might happen that Lorentz violation only stems from terms of dimension greater than 4 in the Lagrangian. Alternatively, if the SME coefficients at leading order are known to have experimental bounds at levels suppressed by two powers of the Planck scale, then it becomes appropriate to look for signals of Lorentz violation at subleading order as well. For these reasons, limits of the SME that contain higher-dimension nonrenormalizable terms are of interest.

In some experiments, particular types of particle behaviors play a major role in attaining sensitivity to Lorentz violation. Examples include spin-precession effects, interference, or flavor-changing oscillations. In these situations it can be advantageous to build specific types of models out of subsets of the SME. Such models can then be used as frameworks for phenomenology. It is also useful to consider complementary tests of Lorentz violation when different experiments only have sensitivity to combinations of SME experiments. An example of this involves CPT tests. These experiments with particles and antiparticles are typically sensitive at leading order to combinations of the CPT-odd terms in the SME. At the same time, different experiments with the particles alone might have leading-order sensitivity to different combinations of both CPT-even and CPT-odd terms. However, by analyzing both sets of experiments in terms of SME coefficients in a complementary manner, it becomes possible to place more stringent bounds on individual types of Lorentz violation.

23.2.1 Constructing SME

The SME contains the SM, a gravity sector, and all possible observer-independent interactions of these conventional fields with fixed Lorentz-violating backgrounds, which are referred to as SME coefficients. As is typically done in field theory, the SME can be constructed in terms of a Lagrangian. The equations of motion are then obtained by variations of the action with respect to the fundamental fields in the theory. The SME Lagrangian has three primary sectors, including one for the SM, one for gravity, and a Lorentz-violating sector,

$$\mathcal{L}_{\text{SME}} = \mathcal{L}_{\text{SM}} + \mathcal{L}_{\text{GRAV}} + \mathcal{L}_{\text{LV}} . \qquad (23.13)$$

The full SME with gravity is defined using a vierbein formalism. This permits a natural distinction between the spacetime manifold and local Lorentz frames.

The observer independence of the SME requires that all of the terms in the Lagrangian be observer scalars under both general coordinate transformations and local Lorentz transformations. This means that every spacetime index and every local Lorentz index must be fully contracted in the Lagrangian.

The SME is not invariant under particle diffeomorphisms and particle local Lorentz transformations. The four infinitesimal parameters ξ^μ comprise the diffeomorphism degrees of freedom, while the six infinitesimal parameters $\epsilon_{ab} = -\epsilon_{ba}$ carry the six Lorentz degrees of freedom. In total, there are ten relevant spacetime symmetries. Violation of these symmetries occurs when an interaction term in the Lagrangian contains SME coefficients that remain fixed under a particle local Lorentz transformation or diffeomorphism.

23.2.2 Minimal SME

Since the SM works remarkably well to describe nongravitational particle interactions at accessible energies, it makes sense initially to construct a minimal extension beyond the SM that contains only those interactions for which experiments are likely to have the greatest sensitivity. These are the interactions that break LLI while maintaining all of the other desirable features of the SM, such as gauge invariance and renormalizability. The mSME is the restriction of the SME to these power-

counting renormalizable and gauge-invariant terms in the absence of gravity.

The mSME Lagrangian can be separated into Lorentz-invariant and Lorentz and CPT-violating parts:

$$\mathcal{L}_{\text{SME,min}} = \mathcal{L}_{\text{SM}} + \mathcal{L}_{\text{LV,min}} . \tag{23.14}$$

The Lorentz-invariant sector is identified with the usual Lagrangian for the minimal SM. The Lorentz-violating Lagrangian is the restriction to terms of mass dimension 3 and 4 in $\mathcal{L}_{\text{LV}}$ that maintain $SU(3)\times SU(2)\times U(1)$ gauge symmetry.

The first component, $\mathcal{L}_{\text{SM}}$, describes the usual interactions for the strong and electroweak interactions. The matter fields consist of three generations of quarks and leptons. These interact through exchange of gauge fields. A Higgs sector is needed to provide mass terms for the W and Z bosons in the weak interactions through the Higgs mechanism, and Yukawa couplings are needed to give the quarks and leptons mass terms as well. The Lagrangian $\mathcal{L}_{\text{SM}}$ can be split into five parts corresponding to these different sectors:

$$\begin{aligned}\mathcal{L}_{\text{SM}} = \mathcal{L}_{\text{lepton}} + \mathcal{L}_{\text{quark}} + \mathcal{L}_{\text{Yukawa}} \\ + \mathcal{L}_{\text{Higgs}} + \mathcal{L}_{\text{gauge}} .\end{aligned} \tag{23.15}$$

For illustration purposes, the form of the terms for the lepton sector are given here:

$$\mathcal{L}_{\text{lepton}} = \tfrac{1}{2}\mathrm{i}\overline{L}_A\gamma^\mu \overleftrightarrow{D}_\mu L_A + \tfrac{1}{2}\mathrm{i}\overline{R}_A\gamma^\mu \overleftrightarrow{D}_\mu R_A . \tag{23.16}$$

In this notation, the left-handed and right-handed lepton multiplets are denoted as

$$L_A = \begin{pmatrix} \nu_A \\ l_A \end{pmatrix}_{\mathrm{L}} , \quad R_A = (l_A)_{\mathrm{R}} . \tag{23.17}$$

The index $A = 1,2,3$ labels the three flavors, with $l_A = (e, \mu, \tau)$ denoting the electron, muon, and tau particles, and $\nu_A = (\nu_e, \nu_\mu, \nu_\tau)$ labeling the three corresponding neutrinos. The gauge-covariant derivative is denoted D_μ, and the notation

$$A \overleftrightarrow{\partial_\mu} B \equiv A\partial_\mu B - (\partial_\mu A)B$$

is adopted. For the remaining terms in $\mathcal{L}_{\text{SM}}$, see [23.2, 3].

The Lorentz and CPT-violating part of the Lagrangian $\mathcal{L}_{\text{LV,min}}$ can also be written as a sum of terms distinguishing the contributions from the lepton, quark, Yukawa, Higgs, and gauge sectors. These partial Lagrangians can be further separated into CPT-even and CPT-odd parts. Each of these terms consists of contractions of the SM fields with the SME coefficients.

To illustrate for the lepton sector, the Lorentz-violating terms are:

$$\begin{aligned}\mathcal{L}_{\text{lepton}}^{\text{CPT-even}} &= \tfrac{1}{2}\mathrm{i}(c_{\mathrm{L}})_{\mu\nu AB}\overline{L}_A\gamma^\mu \overleftrightarrow{D^\nu} L_B \\ &\quad + \tfrac{1}{2}\mathrm{i}(c_{\mathrm{R}})_{\mu\nu AB}\overline{R}_A\gamma^\mu \overleftrightarrow{D^\nu} R_B ,\end{aligned} \tag{23.18}$$

$$\begin{aligned}\mathcal{L}_{\text{lepton}}^{\text{CPT-odd}} &= -(a_{\mathrm{L}})_{\mu AB}\overline{L}_A\gamma^\mu L_B \\ &\quad - (a_{\mathrm{R}})_{\mu AB}\overline{R}_A\gamma^\mu R_B .\end{aligned} \tag{23.19}$$

In these expressions, the SME coefficients a_μ have dimensions of mass, while $c_{\mu\nu}$ are dimensionless and traceless. It is these quantities that act as fixed background fields under particle Lorentz transformations and induce the breaking of Lorentz symmetry.

23.2.3 QED Extension

The QED limit of the SME is useful for specific applications involving charged particle and photon interactions. It contains the leading-order Lorentz- and CPT-violating terms that maintain $U(1)$ gauge invariance. For a single Dirac fermion ψ of mass m the Lagrangian is $\mathcal{L}_{\text{QED, min}} = \mathcal{L}_{\text{fermion}} + \mathcal{L}_{\text{photon}}$. The fermion-sector piece can be written as

$$\mathcal{L}_{\text{fermion}} = \tfrac{1}{2}\mathrm{i}\overline{\psi}\Gamma^\mu \overleftrightarrow{D_\mu}\, \psi - \overline{\psi}M\psi , \tag{23.20}$$

where the gauge-covariant derivative is $D_\mu = \partial_\mu + \mathrm{i}qA_\mu$ and Γ^ν and M are defined by

$$\begin{aligned}\Gamma^\nu &= \gamma^\nu + c^{\mu\nu}\gamma_\mu + d^{\mu\nu}\gamma_5\gamma_\mu + e^\nu \\ &\quad + \mathrm{i}f^\nu\gamma_5 + \tfrac{1}{2}g^{\lambda\mu\nu}\sigma_{\lambda\mu} ,\end{aligned} \tag{23.21}$$

$$M = m + a_\mu\gamma^\mu + b_\mu\gamma_5\gamma^\mu + \tfrac{1}{2}H^{\mu\nu}\sigma_{\mu\nu} . \tag{23.22}$$

These equations contain the usual QED terms for a single fermion. The nonstandard terms violate Lorentz symmetry, and most have analogs in mSME. However, the dimensionless coefficients e^ν, f^ν, $g^{\lambda\mu\nu}$ have no analogue in mSME because they are incompatible with $SU(2)\times U(1)$ symmetry. They are included in the minimal QED extension because they are compatible with $U(1)$ invariance and could emerge from terms in the effective action involving the Higgs field.

The Lagrangian in the photon sector is

$$\mathcal{L}_{\text{photon}} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} - \frac{1}{4}(k_F)_{\kappa\lambda\mu\nu}F^{\kappa\lambda}F^{\mu\nu} + \tfrac{1}{2}(k_{AF})^{\kappa}\epsilon_{\kappa\lambda\mu\nu}A^{\lambda}F^{\mu\nu} . \quad (23.23)$$

For simplicity here, any total derivative terms are neglected, as is a possible term of the form $(k_A)_\kappa A^\kappa$. Some discussion of the latter can be found in [23.2, 3].

In these expressions, the terms with coefficients a_μ, b_μ, e_μ, f_μ, $g_{\lambda\mu\nu}$, and $(k_{AF})_\mu$ are odd under CPT, while those with $H_{\mu\nu}$, $c_{\mu\nu}$, $d_{\mu\nu}$, and $(k_F)_{\kappa\lambda\mu\nu}$ preserve CPT. All ten terms break Lorentz symmetry. Typically, experiments can have different sensitivities to different types of Lorentz violation and can involve different particle species. For this reason, superscript labels are added to the SME coefficients in the fermion sector to denote the particle species. Lagrangian terms of the same form are expected to describe protons and neutrons in QED systems as well, but there the SME coefficients represent composites stemming from quark and gluon interactions.

23.2.4 Extensions in Quantum Mechanics

Many of the sharpest tests of Lorentz symmetry are conducted in high-precision particle and atomic experiments. Typically, static electric and magnetic fields are used in these experiments to trap or control charged particles, while the frequencies of particle transitions between different energy levels are measured with exceptional sensitivity. The electric and magnetic fields can also be manipulated to allow switching between particles and antiparticles, thereby permitting tests of CPT. The leading-order shifts in the standard (Lorentz- and CPT-preserving) energy levels are due to the effects of the small quantities a_μ, b_μ, $H_{\mu\nu}$, $c_{\mu\nu}$, $d_{\mu\nu}$, e_μ, f_μ, and $g_{\lambda\mu\nu}$. Lorentz violation stemming from couplings to the photon coefficients $(k_{AF})_\mu$ and $(k_F)_{\kappa\lambda\mu\nu}$ enters only at subleading order for these types of measurements.

It is often sufficient in calculations describing these systems to work at the level of relativistic quantum mechanics using a modified Dirac equation. It is obeyed by a four-component spinor field ψ describing a particle with charge q and mass m. Calculation of leading-order energy shifts can be carried out most readily within a perturbative framework. To do so requires extracting a suitable Dirac Hamiltonian from the Lagrangian. However, the appearance of time-derivative couplings in the modified Dirac equation means that the standard procedure for obtaining the Dirac Hamiltonian fails to produce a Hermitian quantum mechanical operator generating time translations on the wave function. This technical difficulty can be overcome by performing a field redefinition at the Lagrangian level, chosen to eliminate the additional time derivatives. Rewriting the Lagrangian in terms of the new field χ does not affect the physics. However, the modified Dirac wave function corresponding to χ does have conventional time evolution.

The rewritten Dirac equation takes the form

$$\mathrm{i}\partial_0\chi = \hat{H}\chi , \quad (23.24)$$

with

$$\hat{H} = \hat{H}_0 + \hat{H}_{\text{pert}} . \quad (23.25)$$

In this notation, $\hat{H}_0$ is a conventional Dirac Hamiltonian representing a charged particle in the absence of Lorentz- and CPT-violating perturbations. The perturbative Hamiltonian $\hat{H}_{\text{pert}}$ for the particle is linear in the SME coefficients. The static electromagnetic fields enter in the perturbative treatment at leading order only through the dependence of the gauge-covariant derivatives on the background potential A_μ.

In many experiments, energies are probed only at extremely low energy, where an expansion of the Hamiltonian in a nonrelativistic limit is appropriate. This can be implemented following a Foldy–Wouthuysen approach [23.51]. The resulting nonrelativistic perturbative Hamiltonian can be written in terms of the three-momentum of the particle p_j and the usual Pauli matrices σ^j obeying $[\sigma^j, \sigma^k] = 2\mathrm{i}\varepsilon_{jkl}\sigma^l$. The leading-order terms are

$$\begin{aligned} H_{\text{nonrel}} \simeq{}& m + \frac{p^2}{2m} \\ &+ ((a_0) - mc_{00}) + \\ &+ \left(-b_j + md_{j0}\tfrac{1}{2}\epsilon_{jkl}H_{kl}\right)\sigma^j \\ &+ [-a_j + m(c_{oj} + c_{jo})]\frac{p_j}{m} + \cdots . \end{aligned} \quad (23.26)$$

In nonrelativistic experiments with ordinary matter the primary sensitivity will be to particular combinations of SME coefficients appearing in these terms. Subleading contributions can be calculated from expectation values of the terms involving factors of p_j, where the momentum is treated as a quantum-mechanical operator.

For experiments designed to test CPT, which involves measurements of both particles and antiparticles,

Table 23.1 Transformation properties of dominant SME terms in the matter QED limit under the discrete symmetries C, P, T and their combinations

SME Coeff.	C	P	T	CT	CP	TP	CPT
a_0	−	+	+	−	−	+	−
a_j	−	−	−	+	+	+	−
b_0	+	−	+	+	−	−	−
b_j	+	+	−	−	+	−	−
H_{0j}	−	−	+	−	+	−	+
H_{jk}	−	+	−	+	−	−	+
c_{00}	+	+	+	+	+	+	+
c_{0j}	+	−	−	−	−	+	+
c_{j0}	+	−	−	−	−	+	+
c_{jk}	+	+	+	+	+	+	+
d_{00}	−	−	+	−	+	−	+
d_{0j}	−	+	−	+	−	−	+
d_{j0}	−	+	−	+	−	−	+
d_{jk}	−	−	+	−	+	−	+

the Dirac Hamiltonian for the antiparticle must also be obtained. This is accomplished using charge conjugation. The modified Dirac equation for the antiparticle differs from that of the particle by the sign of the charge q and in the sign of any SME coefficients that are odd under charge conjugation. See Table 23.1 for a list of transformation properties for some of the dominant terms in the QED limit of the mSME.

All of the expressions in the quantum-mechanical limits depend explicitly on the spatial components j, k, l of the SME coefficients and on the components of various physical quantities, such as the particle momenta and the potential A_μ. These components are defined with respect to a laboratory frame that must be chosen with a particular orientation. In laboratory frames fixed with respect to the surface of the Earth, the $j = 3$ (or z direction) is usually chosen as the relevant quantization axis, typically corresponding to the direction of a static magnetic field. Alternatively, if a rotation device is used on Earth's surface, such as a turntable, its orientation can be chosen as the $j = 3$ direction. In a moving lab, such as in a satellite orbiting the Earth, a standard configuration defines the $j = 3$ direction along the satellite velocity with respect to Earth, with the $j = 1$ direction pointing toward Earth and the $j = 2$ direction completing the right-handed system. In certain situations, Earth-based experiments may choose to use a satellite-based configuration as well, where the velocity of motion is due to the rotation of the Earth about its axis. The objective in this case is to take boost effects into account on the surface of the Earth, as is done in satellite experiments. Ultimately, no matter which of these alignments is chosen for the lab frame directions labeled by j, the laboratory axes must be referenced to a nonrotating basis that can serve as a standard, since it is only with such a standard basis that comparisons across different experiments can be made. Bounds on components of SME coefficients in the lab frame must, therefore, be mapped into bounds on their components with respect to the standard reference frame.

For the standard reference frame, there are a number of different choices that could be made. Examples include reference frames attached to the centers of mass of the Earth, the Sun, the Milky Way galaxy, and the cosmic microwave background radiation (CMBR). With the exception of the Earth, each is approximately inertial over thousands of years. Typically in experiments, a Sun-centered celestial equatorial frame is chosen as the standard reference frame. It is used as the basis for reporting sensitivities to Lorentz violation. In certain limits, e.g., over short time scales where effects of boosts can be ignored, the spatial Sun-centered spatial components reduce to corresponding values in an Earth-based frame. Similarly, observer transformations from the Sun frame to a galaxy-based or CMBR-based frame can be made if bounds are desired with respect to these frames.

23.2.5 Gravity Sector

The gravity sector of the SME uses a vierbein formalism, which gives the theory a close parallel to gauge theory. Lorentz breaking occurs due to the presence of SME coefficients, which remain fixed under particle Lorentz transformations in a local frame. In this case, the SME coefficients carry Latin indices, e.g., b_a for a vector, with respect to the local basis set. The conversion to spacetime coordinates is implemented by the vierbein, giving, e.g., $b_\mu = e_\mu{}^a b_a$. The Lagrangian can then be written in terms of fields and SME coefficients defined on the spacetime manifold. A natural (although not required) assumption is that the SME coefficients are smooth functions over the manifold. It is not necessary to require that they be covariantly constant. In fact, defining covariantly constant tensors over a manifold places stringent topological constraints on the geometry. One simplifying assumption, which could occur naturally in the context of spontaneous Lorentz breaking, is to assume that the SME coeffi-

cients are constants in the local frame. However, again, this is not a requirement in the formulation of the SME theory.

To construct the mSME including gravity [23.5], the first step is to incorporate gravitational fields into the usual SM. This is done by rewriting all of the terms in the SM Lagrangian with fields and gamma matrices defined with respect to the local frame (using Latin indices). The vierbein is then used to convert these terms over to the spacetime manifold. Factors of the determinant of the vierbein e are included as well, so that integration of the Lagrangian density (giving the action) is covariant. Derivatives are understood as well to be both spacetime and gauge-covariant. With these changes, (23.16), for example, becomes

$$\mathcal{L}_{\text{lepton}} = \tfrac{1}{2}\mathrm{i}ee^{\mu}{}_{a}\overline{L}_A\gamma^a \overleftrightarrow{D}_\mu L_A + \tfrac{1}{2}\mathrm{i}ee^{\mu}{}_{a}\overline{R}_A\gamma^a \overleftrightarrow{D}_\mu R_A \,. \quad (23.27)$$

The other terms for the quark, Yukawa, Higgs, and gauge sectors follow a similar pattern.

The Lorentz-violating SME terms constructed from SM fields are obtained in a similar way. The various particle sectors can again be divided between CPT odd and even contributions. Each of the terms in the Lagrangian is then written using local indices and vierbeins, which convert the equations over to the spacetime manifold. As an example, (23.18) becomes

$$\mathcal{L}^{\text{CPT-even}}_{\text{lepton}} = -\tfrac{1}{2}\mathrm{i}(c_{\mathrm{L}})_{\mu\nu AB}ee^{\mu}{}_{a}\overline{L}_A\gamma^a \overleftrightarrow{D^\nu} L_B - \tfrac{1}{2}\mathrm{i}(c_{\mathrm{R}})_{\mu\nu AB}ee^{\mu}{}_{a}\overline{R}_A\gamma^a \overleftrightarrow{D^\nu} R_B \,. \quad (23.28)$$

The remaining equations follow the same pattern.

The pure-gravity sector of the mSME consists of a Lorentz-invariant gravity sector and a Lorentz-violating sector. The Lorentz-invariant Lagrangian consists of terms that are products of the gravitational fields. In the general case, this includes terms constructed from curvature, torsion, and covariant derivatives. Einstein's gravity (with or without a cosmological term) would be a special case in this sector.

The Lorentz-violating Lagrangian terms in the gravity sector of the mSME are constructed by combining the SME coefficients with gravitational field operators to produce an observer scalar under local Lorentz transformations and general coordinate transformations. These consist of products of the vierbein, the spin connection, and their derivatives, but for simplicity they can be written in terms of the curvature, the torsion $T_{\lambda\mu\nu}$, and covariant derivatives. A minimal case (up to dimension four) has the form:

$$\mathcal{L}^{\text{LV}}_{e,\omega} = e(k_T)^{\lambda\mu\nu}T_{\lambda\mu\nu} + e(k_{\mathrm{R}})^{\kappa\lambda\mu\nu}R_{\kappa\lambda\mu\nu} + e(k_{TT})^{\alpha\beta\gamma\lambda\mu\nu}T_{\alpha\beta\gamma}T_{\lambda\mu\nu} + e(k_{DT})^{\kappa\lambda\mu\nu}D_\kappa T_{\lambda\mu\nu} \,. \quad (23.29)$$

The SME coefficients in this expression have the symmetries of the associated Lorentz-violating operators that they multiply.

The Lorentz-violating sector introduces additional gravitational couplings that can have phenomenological consequences, including effects on cosmology, black holes, gravitational radiation, and post-Newtonian physics. As a starting point for a phenomenological investigation of the gravitational consequences of Lorentz violation, it is useful to write down the Riemannian limit of the mSME gravity sector. It is given as [23.5]

$$S_{e,\omega,\Lambda} = \frac{1}{2\kappa}\int \mathrm{d}^4x \Big[e(1-u)R - 2e\Lambda + es^{\mu\nu}R_{\mu\nu} + et^{\kappa\lambda\mu\nu}R_{\kappa\lambda\mu\nu}\Big] \,. \quad (23.30)$$

The SME coefficient $(k_{\mathrm{R}})^{\kappa\lambda\mu\nu}$ has been expanded into coefficients $s^{\mu\nu}$, $t^{\kappa\lambda\mu\nu}$, u that distinguish the effects involving the Riemann, Ricci, and scalar curvatures. The coefficients $s^{\mu\nu}$ have the symmetries of the Ricci tensor, while $t^{\kappa\lambda\mu\nu}$ has those of the Riemann tensor. Taking tracelessness conditions into account, there are 19 independent components.

23.2.6 Spontaneous Lorentz Violation

There are a number of theoretical issues concerning Lorentz violation that can be examined using the SME. One concerns the nature of the symmetry breaking and how that affects the interpretation of the SME coefficients. These coefficients, e.g., b_μ, for the case of a vector, couple to the SM and gravitational fields as fixed backgrounds. For the case of a single fermion field, ψ, in special relativity, the coupling has the form, $b_\mu\overline{\psi}\gamma^5\gamma^\mu\psi$. If this is the only term in the SME Lagrangian containing the coefficient b_μ, then the

symmetry breaking is said to be explicit. Essentially the coefficient b_μ appears in the effective field theory without any underlying dynamics. However, it is also possible for the SME coefficients to arise through a process of spontaneous symmetry breaking. In this case, the SME coefficients are interpreted as vevs of a dynamical tensor field. For example, for a vector B_μ, the SME coefficient would arise as a vev, $\langle B_\mu \rangle = b_\mu$. The vev acts as a fixed background field that spontaneously breaks Lorentz symmetry, but the vector B_μ remains fully dynamical.

The process of spontaneous symmetry breaking is important in particle physics. For example, in the electroweak theory, the scalar Higgs field ϕ acquires a nonzero vev, $\langle \phi \rangle \neq 0$, that spontaneously breaks the local $SU(2) \times U(1)$ gauge symmetry. For a scalar field, there is no associated breaking of Lorentz symmetry because the scalar vev is invariant under Lorentz transformations. However, the SME coefficients have tensor indices. When these occur as nonzero vevs, Lorentz symmetry is said to be spontaneously broken.

The standard construction of the SME does not make a distinction between whether the breaking of Lorentz symmetry is explicit or spontaneous. Both types of symmetry breaking can be accommodated, and both are useful to consider for phenomenological investigations of Lorentz violation. However, when the gravitational sector of the SME is included, which brings more geometrical considerations into play, it becomes important to distinguish these types of symmetry breaking.

For the case of explicit Lorentz violation, it has been shown that inconsistencies arise between geometrical constraints (e.g., Bianchi identities) and conditions stemming from the equations of motion. This was proved by *Kostelecký* in a no-go theorem [23.5]. However, the no-go theorem is evaded if the symmetry breaking is spontaneous. The crux of the difference has to do with the fact that if the Lorentz breaking is spontaneous, then all of the SME coefficients must be treated as dynamical fields in field variations.

Because of this, it is often assumed that the SME coefficients are, indeed, vevs of dynamical fields that have undergone a process of spontaneous Lorentz breaking. Note, however, that if the vevs are associated with very high energy scales, then in low-energy tests of Lorentz violation, they will still act primarily as fixed background fields, and their dynamics at higher energies will not be relevant. It is for this reason that the form of the SME or mSME used by most experimentalists is the same as if the symmetry breaking were explicit. For purposes of phenomenology, the distinction between explicit and spontaneous Lorentz breaking is not crucial. For the case of explicit breaking, it may be that a different type of geometry is relevant, known as a Riemann–Finsler geometry (for a review, see [23.52]). The SME with explicit breaking has been shown to be linked to Riemann–Finsler geometry [23.53, 54].

It is certainly the case that spontaneous symmetry breaking is a very elegant form of symmetry breaking. This is because when a symmetry is spontaneously broken, it still holds dynamically. However, the vacuum solution for the theory does not obey the symmetry. What is often done is that a field redefinition is performed that resets the vacuum values to zero. In this case, in terms of the new set of fields, the symmetry becomes hidden at the level of the equations of motion. It is for this reason that spontaneous symmetry breaking is also referred to as hidden symmetry.

From a theoretical point of view, there are well-known consequences when a symmetry is spontaneously broken. For example, when a global continuous symmetry is spontaneously broken, it has been shown that massless fields, called Nambu–Goldstone (NG) fields appear [23.55–57]. On the other hand, if the symmetry is local, as in the case of the electroweak model, then a Higgs mechanism can occur [23.58–60]. In this case, the would-be NG modes are reinterpreted in a way that results in the gauge fields acquiring a mass. This is what happens in the electroweak model, and as a result the W and Z bosons are massive. However, an unbroken local $U(1)$ gauge symmetry allows the photons to remain massless. At the same time, there are excitations of the Higgs scalar field that are also massive. This results in a massive Higgs boson, which has recently been detected at the Large Hadron Collider.

An important theoretical issue to consider is whether these same types of processes can occur when it is Lorentz symmetry that is spontaneously broken. For the case where Lorentz symmetry is global, as in the context of special relativity, the Goldstone theorem would suggest that massless NG modes should appear. If so, they would appear as infinite-range particles and would have implications for phenomenology. The only known massless particles in the SM and GR (assuming neutrinos have mass) are the gauge fields, such as the photon, graviton, and gluons. Thus, it would seem that there are only two possibilities for the NG modes. Either the NG modes are known particles, such as photons or gravitons, or they are unknown fields that have escaped detection. However, if the Lorentz symmetry is local, as in a gravitational theory, then the question of

whether a Higgs mechanism can occur becomes relevant. In this case, the possibility of massive gauge fields arises (massive photons or massive gravity), and the question of whether there are additional massive Higgs fields needs to be addressed as well.

These types of questions have been investigated both in special relativity and in the context of gravity using models that are subsets of the SME. Interestingly, some of these investigations occurred before the process of spontaneous symmetry breaking was fully understood. For example, Dirac worked with a vector model that had a constraint that the norm of the vector be nonzero [23.61]. Nambu later showed that such a model spontaneously breaks Lorentz symmetry [23.62]. Bjorken found a similar model using a composite theory of fermions that collectively have a nonzero vector vev [23.63]. It was conjectured that in these types of models, the NG modes can be interpreted as photons. This raises the interesting possibility that photons are massless because they are the NG modes associated with spontaneous Lorentz breaking, whereas the conventional idea is that photons are massless because of local gauge invariance.

In order to impose a constraint that a vector field has a nonzero vev, the usual process in field theory is to include a potential term that has a minimum when the vector field equals its vev. Theories with a vector field and a potential of this type that induces spontaneous Lorentz violation are known as bumblebee models [23.5, 15, 64–76]. A defining feature of these theories is that they do not have local $U(1)$ gauge invariance. Thus, there is no possibility in these models for photons to arise because of local $U(1)$ gauge symmetry. Recent investigations of bumblebee models have shown that all of the usual processes associated with spontaneous symmetry breaking can occur when the symmetry is Lorentz symmetry. First, however, it was found that there is a link between local Lorentz symmetry and diffeomorphisms. In general, if one of these symmetries is spontaneously broken, then so is the other. For example, if a vector field has a vev b_a in a local Lorentz frame, which spontaneously breaks LLI, then it will also have a vev b_μ in the spacetime frame, which spontaneously breaks diffeomorphisms. (Even for a scalar vev with spacetime dependence this is true, although in this case it is the derivatives of the scalar that spontaneously break the symmetries). What this means is that in the context of a gravitational theory with spontaneous Lorentz breaking there can be up to ten NG modes, six associated with Lorentz breaking, and four associated with diffeomorphism breaking.

If these symmetries are treated analogously to local gauge symmetry using a vierbein formalism, then it is possible to show that the vierbein itself can accommodate all ten NG modes when local Lorentz symmetry and diffeomorphisms are spontaneously broken. It is also possible to investigate whether a Higgs mechanism can occur and whether additional massive Higgs modes can appear. Interestingly, it is found that for a Higgs mechanism to occur the geometry cannot be Riemannian. This is because the gauge fields associated with the local Lorentz symmetry are the spin connection, and in order to have a dynamical spin connection, the theory must include torsion. The geometry must, therefore, be Riemann–Cartan if a Higgs mechanism is to occur. There can also be additional massive Higgs modes that can affect the propagation of metric excitations (or gravitational radiation). It is for this reason that theories of massive gravity often result from the process of spontaneous Lorentz violation. In all of these models, there are stringent conditions that must hold so that unphysical modes do not appear, such as negative energy states or tachyons. These constraints very severely limit the possibilities for making viable models with massive gravitational fields or massive propagating spin connection.

A subset of the bumblebee models in which the kinetic term for the vector field has a Maxwell form, are known as Kostelecký–Samuel (KS) models [23.15]. For these models, it has been shown that in the limit where the massive Higgs modes becomes extremely massive, the solutions for the KS model match those of Einstein–Maxwell theory in a fixed gauge. Thus, the intriguing idea that photons might arise as NG modes in a theory with spontaneous Lorentz breaking still holds even when gravity is included.

It is also possible to consider models with other types of tensor fields that acquire nonzero vacuum values. Some possibilities include theories with a symmetric two-tensor or alternatively an anti-symmetric two-tensor [23.77–79]. Just as with a vector, when Lorentz symmetry in these models is spontaneously broken, NG modes and massive modes can appear. It is useful to study these models to see what the various possibilities are for the NG and massive modes. One interesting case is that of a symmetric two-tensor in a Minkowski background. In this type of model, known as a cardinal model, the NG modes have properties similar to the graviton in GR, but in a fixed gauge. This again raises the intriguing question of whether known massless particles might occur as a result of Lorentz breaking. A related consideration is then whether there exist sig-

natures of the Lorentz breaking that can distinguish KS and cardinal models from conventional physics. These types of phenomenological questions can then be suitably addressed in the context of the SME.

In cosmology, models with spontaneous Lorentz violation have been used to study modifications of gravity that might give rise to effects such as accelerated expansion of the universe or to introduce anisotropic features in the cosmic background radiation. Examples include [23.80–86]. In general, these models, which incorporate vector or tensor fields that spontaneously break Lorentz symmetry, are studied as possible alternative theories of dark energy. While these theories have a number of interesting effects and features, they do not typically give rise to high-precision observational constraints on LLI. For this reason, these models are not considered here, and the reader is referred to the literature.

23.3 Experimental Tests of Lorentz Violation

If Lorentz invariance is not an exact symmetry due to mechanisms occurring in the context of a quantum theory of gravity, then the relevant energy scale is presumably the Planck scale, since this is the scale where gravity meets up with quantum physics. At one time, it was thought unlikely that any physics arising from the Planck scale would be accessible to experimental detection. However, with Lorentz violation, the Planck scale is expected to enter as a suppression factor or inverse power in any corrections to conventional physics. Therefore, instead of needing to accelerate particles to ultra-high energies that are impossible to obtain, one can look at extremely high-precision experiments often at very low energies for signs of Planck-scale physics. In this approach, Lorentz breaking provides an ideal signal of new physics, since nothing in the SM permits violation of Einstein's theory. That is, no conventional process could ever mimic or cover up a genuine signal of Lorentz violation.

The SME serves as a common framework used by experimentalists and theorists to search for signals of Lorentz and CPT violation. Planck-scale sensitivity has been attained to the dominant SME coefficients in a number of experiments involving different particle sectors. These include experiments with mesons, photons, electrons, protons, neutrons, muons, neutrinos, and in the electroweak sector. Each particle sector has unique features, and the experimental methods for testing Lorentz and CPT violation can differ case by case.

In some experiments, leading-order sensitivity to Lorentz and CPT violation exists for more than one particle species at the same time. This is particularly true in atomic experiments where bounds involving all three of the electron, proton, and neutron are often obtained. Likewise, mixtures of flavors in the meson and neutrino sectors can occur naturally. In these cases, the experimental bounds obtained are for combinations of SME coefficients for the different particle sectors. It is, therefore, important to look for complementary sets of bounds obtained from different experiments that can be combined to select out an optimal set of bounds for the individual particle species.

In a similar manner, experiments can have sensitivity to either both CPT-odd and CPT-even forms of Lorentz violation, or alternatively they can probe only the CPT-odd sector. Most bounds obtained typically involve combinations of both CPT-odd and CPT-even SME coefficients. However, experiments designed to test CPT switch between measurements on particles and similar measurements on the corresponding antiparticles. The bounds in this case are only on CPT-odd SME coefficients. For a given particle species, performing both types of experiments provides a natural complementary approach.

Before looking at specific experiments, it is useful to examine some general features that are common to a number of different experiments. For example, in low-energy atomic tests, the sensitivity stems primarily from the ability of these experiments to detect extremely small anomalous energy shifts. In many cases, these energy shifts result in small frequency shifts that can be measured with very high precision. It is not uncommon for an atomic experiment to be able to measure a frequency shift with a precision of 1 mHz or less. Interpreting this as being due to an energy shift expressed in GeV, it corresponds to a sensitivity of approximately 4×10^{-27} GeV. Such a value is well within the range of energy one might associate with suppression factors originating from the Planck scale. While many of the original atomic experiments were designed to measure specific quantities, such as charge-to-mass ratios of particles and antiparticles or differences in g factors, it turns out that it is more effective for these experiments to investigate the lowest attainable energy levels

for possible anomalous shifts associated with Lorentz violation. Many experiments look specifically for sidereal time variations of energy levels of a particle or atom as the Earth moves. These would result from interactions with the fixed Lorentz-violating background fields. Alternatively, experiments designed to test CPT can look for instantaneous differences in the energy levels of a particle (or atom) and its antiparticle (or antiatom).

Another important general consideration is the choice of a standard inertial reference frame [23.87]. Laboratory measurements of Lorentz and CPT symmetry involve components of SME coefficients defined with respect to a local laboratory coordinate system. These components labeled with indices $\{0,j\}$ change as the lab frame moves or rotates with respect to an inertial frame. In order to give measured bounds in a consistent manner, these laboratory bounds must be related to bounds on SME coefficients defined with respect to a standard inertial frame. The usual choice for this frame is a Sun-centered frame that uses celestial equatorial coordinates. Components with respect to the Sun-centered frame are denoted using upper-case letters $J, K, L, \ldots$ that run over four independent directions labeled as $\hat{T}$, $\hat{X}$, $\hat{Y}$, and $\hat{Z}$. The spatial origin of this system is the Sun's center, and the unit vector $\hat{Z}$ points along the Earth's rotation axis, while $\hat{X}$ and $\hat{Y}$ lie in the equatorial plane with $\hat{X}$ pointing towards the vernal equinox in the celestial sphere. The time T is measured by a stationary clock at the origin, with $T=0$ taken as the vernal equinox in the year 2000. The Earth's orbital plane lies at an angle $\eta \simeq 23°$ with respect to the XY plane.

Earth-based experiments sensitive to sidereal time variations are sensitive to a combination of coefficients, which are often denoted collectively using tildes. For example, for electrons, the combination of spatial components in the lab frame

$$\tilde{b}^e_j \equiv b^e_j - m d^e_{j0} - \tfrac{1}{2}\varepsilon_{jkl} H^e_{kl} , \tag{23.31}$$

frequently arises in a number of experiments. These combinations are projected onto the nonrotating frame, where the components are b^e_X, b^e_Y, b^e_Z, etc. Nonrotating frame analogs of the coefficient combinations in (23.31) can be defined as

$$\tilde{b}^e_J \equiv b^e_J - m d^e_{J0} - \tfrac{1}{2}\varepsilon_{JKL} H^e_{KL} , \tag{23.32}$$

where J,K,L label the spatial directions X,Y,Z in the nonrotating frame. Ignoring boost effects, the relation between the laboratory and nonrotating spatial components is

$$\begin{aligned}
\tilde{b}^e_1 &= \tilde{b}^e_X \cos\chi \cos\Omega t \\
&\quad + \tilde{b}^e_Y \cos\chi \sin\Omega t - \tilde{b}^e_Z \sin\chi , \\
\tilde{b}^e_2 &= -\tilde{b}^e_X \sin\Omega t + \tilde{b}^e_Y \cos\Omega t , \\
\tilde{b}^e_3 &= \tilde{b}^e_X \sin\chi \cos\Omega t \\
&\quad + \tilde{b}^e_Y \sin\chi \sin\Omega t + \tilde{b}^e_Z \cos\chi .
\end{aligned} \tag{23.33}$$

The angle χ is between the $j=3$ lab axis and the direction of the Earth's rotation axis along Z. The angular frequency $\Omega \simeq 2\pi/(23\,\text{h}\,56\,\text{m})$ is that corresponding to a sidereal day.

23.3.1 Data Tables

A wide range of particle sectors has been investigated for Lorentz and CPT violation. Many experiments achieve very high sensitivity to Lorentz violation and are able to place stringent bounds on the relevant SME coefficients. The results for these bounds are too extensive to list here. However, a comprehensive summary of Lorentz and CPT tests has been published by *Kostelecký*'s group at Indiana University [23.88]. It is also updated annually in the physics archive.

The data tables in [23.88] provide bounds on Lorentz violation for ordinary matter (electrons, protons, and neutrons), photons, mesons, muons, neutrinos, the electroweak sector, and gravity. Many tests compare particles and antiparticles. Low-energy tests in atomic physics include experiments in Penning traps, comparisons of atomic clocks and masers, experiments with atomic fountains, and experiments with antihydrogen at CERN. Photon tests have been performed using astrophysical and cosmological sources as well as resonant cavities in the microwave and optical regimes. Cosmic rays have been investigated for features associated with Lorentz violation. Experiments with mesons, muons, and neutrinos have used large accelerators at high energies. Experiments are planned or underway on the International Space Station (ISS), in space satellites, or using detectors at the south pole. Experiments with macroscopic torsion pendula take advantage of the alignment of large numbers of electron spins to provide bounds with extremely high sensitivity. To measure boost effects, some experiments collect data over long periods of time to enable the Earth's motion to be included. Other experiments use rotating platforms to gain sensitivity to a wider range of space-time directions.

The extremely tight experimental bounds that have been obtained on the leading-order SME coefficients indicate that if Lorentz or CPT violation does occur in nature, it results in only very small corrections to the SM and GR at ordinary energies. Since an underlying fundamental theory that would permit calculation of these corrections is lacking, at best only order-of-magnitude estimates can be given for the leading-order SME coefficients. One possibility is that the leading-order Lorentz-violating terms in the SME are suppressed by at least one inverse power of the Planck scale. If a ratio is formed with a low-energy scale on the order of 1 GeV with the Planck scale, this results in a suppression factor on the order of 10^{-19}. Interestingly, many of the recent experiments that test Lorentz and CPT symmetry have sensitivities that are comparable to or exceed expected order-of-magnitude values based on this suppression factor. For this reason, it is important as well to search for Lorentz violation stemming from subleading-order terms that are not included in the mSME. A systematic treatment of these higher-dimensional terms in the SME has been developed for certain particle sectors, and bounds on some of these coefficients are included in the data tables as well.

23.3.2 Examples

To highlight some of the Lorentz and CPT tests that have been performed a number of different experimental approaches are described here. In many cases, bounds on a selective subset of SME coefficients are given. For a full list of experiments with published bounds on SME coefficients, the reader is referred to the data tables in [23.88].

- *Penning traps* [23.89–93]: Experiments in Penning traps use electric and magnetic fields to isolate and study individual particles and antiparticles. There are two leading-order signals of Lorentz and CPT violation in the electron sector that have been probed in these experiments. One looks for sidereal time variations in the electron cyclotron and anomaly frequencies. The idea here is that the Lorentz and CPT-violating interactions depend on the orientation of the quantization axis in the laboratory frame, which changes as the Earth turns on its axis. As a result, both the cyclotron and anomaly frequencies have small corrections which cause them to exhibit sidereal time variations. Such a signal can be measured using just electrons. Measured bounds are expressed in terms of components in the nonrotating Sun-centered frame for the combination given in (23.32). Their numerical values are on the order of $|\tilde{b}^e_J| \lesssim 10^{-24}$ GeV for $J = X, Y$. The second type of test in a Penning trap is a traditional CPT test that compares electrons and positrons directly. It looks for an instantaneous difference in their anomaly frequencies. Leading-order sensitivity in this case involves only the CPT-odd coefficient b^e_3 (with no tilde), which is the component of b^e_μ along the quantization axis in the laboratory frame. The bound obtained for $|b^e_3|$ is on the order of 10^{-25} GeV.
- *Torsion pendulum* [23.94–97]: Experiments using a spin-polarized torsion pendulum are able to achieve very high sensitivity to Lorentz violation because the torsion pendulum has a huge number of aligned electron spins but a negligible magnetic field. For example, a pendulum at the University of Washington is built out of a stack of toroidal magnets, which has a net electron spin $S \simeq 10^{23}$. The apparatus is suspended on a rotating turntable and the time variations of the twisting pendulum are measured. An analysis of this system shows that in addition to a signal having the period of the rotating turntable, the effects due to Lorentz and CPT violation also cause additional time variations with a sidereal period caused by the rotation of the Earth. Sensitivity to the electron coefficients has been obtained at the levels of $|\tilde{b}^e_J| \lesssim 10^{-31}$ GeV for $J = X, Y$ and $|\tilde{b}^e_Z| \lesssim 10^{-30}$ GeV. By analyzing data over the course of a year, taking the Earth's motion around the Sun into account, sensitivity to Lorentz-boost violating coefficients has been attained as well. This involves a suppression by $v/c \simeq 10^{-4}$, where v is the velocity of the Earth around the Sun. The bound on the time-like combination of coefficients is $\tilde{b}^e_T \lesssim 10^{-27}$ GeV.
- *Clock-comparison tests* [23.87, 98–106]: Many of the sharpest Lorentz bounds for protons and neutrons stem from atomic clock-comparison experiments. These involve making high-precision comparisons of atomic clock signals as the Earth rotates. The clock frequencies are typically hyperfine or Zeeman transitions. Experiments have used hydrogen masers and two-species noble-gas masers to achieve the highest sensitivities to Lorentz violation. For example, a recent experiment with a K-He^3 comagnetometer obtained a bound in the neutron sector equal to $|\tilde{b}^n_J| \lesssim 10^{-33}$ GeV for $J = X, Y$ [23.107]. Experiments with hydrogen masers attain exceptionally sharp sensitivity to Lorentz

and CPT violation in the electron and proton sectors. These experiments use a double-resonance technique that does not depend on there being a field-independent point for the transition. The sensitivity for the proton attained in these experiments is $|\tilde{b}_J^p| \lesssim 10^{-27}$ GeV. Due to the simplicity of hydrogen, this is an extremely clean bound and is one of the more stringent tests for the proton. Clock-comparison experiments performed in space would have several advantages over traditional ground-based experiments. For example, a clock-comparison experiment conducted aboard the ISS would be in a laboratory frame that is both rotating and boosted. It would, therefore, immediately gain sensitivity to a wide range of SME coefficients that are currently untested [23.108, 109]. A European mission is planned for the ISS, which will compare atomic clocks and H masers.

- *Antihydrogen* [23.98, 110, 111]: The ALPHA and ATRAP experiments underway at CERN are designed to produce antihydrogen and to do high-precision spectroscopy on it. One objective is to make high-precision spectroscopic measurements of the 1S–2S transitions in hydrogen and antihydrogen. These are forbidden (two-photon) transitions that have a relative linewidth of approximately 10^{-15}. The ultimate goal is to measure the line center of this transition to a part in 10^3 yielding a frequency comparison between hydrogen and antihydrogen at a level of 10^{-18}. An alternative to 1S–2S transitions is to consider the sensitivity to Lorentz violation in ground-state Zeeman hyperfine transitions. It is found that there are leading-order corrections in these levels in both hydrogen and antihydrogen. Comparing these measurements for hydrogen and antihydrogen will provide a direct CPT test.
- *Photon tests* [23.112–124]: The relevant leading-order terms for the photon sector in the SME are the k_{AF} and k_F terms in (23.23). For the coefficient k_{AF}, which is odd under CPT, it is found theoretically that this term leads to negative energy contributions and is a potential source of instability in the theory unless it is set to zero [23.125]. In addition, very stringent experimental constraints that come from studying the polarization of radiation from distant radio galaxies also exist and are consistent with $k_{AF} \approx 0$. The terms with coefficients k_F are even under CPT and provide positive energy contributions. There are 19 independent components in the k_F coefficients. Ten of these lead to birefrigence of light. Bounds on these coefficients of order 10^{-32} have been obtained from spectropolarimetry of light from distant galaxies. The remaining nine coefficients have been bounded in a series of laboratory photon experiments. These include experiments using optical and microwave cavities, an Ives–Stilwell experiment, and experiments using rotating platforms. Sensitivities ranging from 10^{-9} up to 10^{-17} have been attained for these coefficients.
- *Cosmic rays* [23.126–128]: Cosmic rays provide the highest-energy particles available experimentally and can be used to study LLI. In the presence of Lorentz violation, the maximal attainable velocity for a cosmic ray in vacuum can be different from the speed of light by a small amount. In principle, it can even exceed the speed of light. Effects of this difference include the possibility of photon decay into electron–positron pairs or vacuum Cerenkov radiation by ultra-high energy electrons, both of which are forbidden in the SM. Another effect is the prediction in the context of the SM and special relativity that an upper energy limit known as the Greisen–Zatsepin–Kuzmin, or GZK limit [23.129, 130], should hold for cosmic rays emitted from distant sources. This theoretical limit is set by interactions with the cosmic microwave background radiation over long distances. However, in the presence of Lorentz violation, it is possible for high-energy cosmic rays from distant sources to exceed the GZK limit. This, therefore, provides an opportunity for testing LLI and obtaining bounds on the relevant SME coefficients. Recent experiments at the high resolution fly's eye (HiRes) and Pierre Auger Observatory have searched for ultra-high energy cosmic rays above the GZK limit, and their results appear to confirm the existence of the GZK cutoff.
- *Meson tests* [23.131–137]: Experiments involving neutral meson oscillations provide very sharp tests of Lorentz and CPT symmetry. These investigations attain high sensitivity to the CPT-odd a_μ coefficients in the SME for the K, D, B_d, and B_s meson systems. The time evolution of a meson and its antimeson can be described by an effective Hamiltonian in a description based on the Schrödinger equation. The dominant Lorentz and CPT-violating contributions to the effective Hamiltonian can be calculated as expectation values of interaction terms in the SME. The results depend on the velocity of the meson with respect to the laboratory frame and the combinations of SME coefficients Δa_μ,

which vary with sidereal time as the Earth rotates. Recent analyses have attained bounds on the order of 10^{-21} GeV for neutral kaons, 10^{-15} GeV in the D system, 10^{-14} GeV for B_d oscillations, and 10^{-12} GeV for B_s oscillations.

- *Muon tests* [23.138–140]: Lorentz and CPT tests with muons involve second-generation leptons and are independent of the tests involving electrons. Several different types of experiments with muons have been conducted, including muonium experiments and $g-2$ experiments with muons. In muonium, experiments measuring the frequencies of ground-state Zeeman hyperfine transitions in a strong magnetic field have the greatest sensitivity to Lorentz and CPT violation. A recent analysis has searched for sidereal time variations in these transitions. A bound on SME coefficients, $|\tilde{b}_J^{\mu}|$, has been obtained at a level of 10^{-23} GeV. In relativistic $g-2$ experiments using positive and negative muons bounds on Lorentz-violation SME coefficients have been obtained at a level of 10^{-24} GeV.
- *Collider tests* [23.141–144]: High energy experiments at colliders provide opportunities for testing Lorentz and CPT violation in the QED and quark sectors. Sensitivity for Lorentz violation in cross sections and decay rates has been investigated in electron-positron scattering. Effects include variations in observed cross sections with periodicities controlled by Earth's sidereal rotation frequency. In a recent experiment using the D0 detector at the Fermilab Tevatron Collider, a search for violation of Lorentz invariance in the top quark-antiquark production cross section was carried out, and bounds on SME coefficients for the top quark were obtained.
- *Neutrino tests* [23.145–155]: The experimental observation that neutrinos change flavor when they propagate through space cannot be explained by the SM. The conventional explanation for these neutrino oscillations is that the particles have very small masses. However, at the same time, the high-precision sensitivity of neutrino oscillation experiments, stemming from their interferometric nature, offers possibilities for a range of new tests of LLI. The neutrino sector of the mSME contains Lorentz-violating interactions for left-handed neutrinos and right-handed antineutrinos. For the left-handed neutrinos, sensitivity at leading order is to the SME coefficients $(a_L)^{\mu}$ and $(c_L)^{\mu\nu}$. The resulting signals include ones with the usual L/E dependence, where E is the energy and L is the oscillation length or baseline of the experiment. However, with Lorentz violation other dependences, such as ones with L or LE are possible as well. These lead to unique signatures of Lorentz violation that can occur in neutrino experiments. These include oscillation, time of flight, and threshold effects. For example, it has been shown that a Lorentz-violating seesaw mechanism can occur, which allows for oscillatory behavior even in the absence of mass. The coefficients for Lorentz violation can also couple to the four-momentum of the neutrino. In terrestrial experiments, the direction of the neutrino beam changes as the Earth rotates, which leads to sidereal time variations in the oscillation data when LLI is broken. The mSME has been applied to a number of neutrino experiments, including both short-baseline and long-baseline experiments. An extensive list of bounds on SME coefficients in the neutrino sector are given in the data tables [23.88]. For the coefficients $(a_L)^{\mu}$, bounds at the level of 10^{-20}–10^{-23} GeV have been obtained, while for the $(c_L)^{\mu\nu}$ coefficients, the sensitivity ranges from 10^{-17} to 10^{-27}.
- *Gravity tests* [23.156–163]: Lorentz violation in the gravity sector stems from both matter–gravity couplings and pure gravity couplings. In some cases, the matter–gravity couplings can lead to sensitivity to forms of Lorentz violation that would otherwise go undetected in the absence of gravity. The leading-order SME terms for both these sectors in a linearized gravity regime involve expectation values denoted as $\bar{a}_{\mu}$, $\bar{c}_{\mu\nu}$ and $\bar{s}_{\mu\nu}$. At leading order, matter–gravity tests are sensitive to $\bar{a}_{\mu}$ and $\bar{c}_{\mu\nu}$, while pure gravity tests are sensitive to $\bar{s}_{\mu\nu}$. The matter–gravity tests include gravimeter, atom interferometry, and weak equivalence principle experiments. Bounds on $\bar{a}_{\mu}$ have been obtained at levels of 10^{-6}–10^{-11} GeV and on $\bar{c}_{\mu\nu}$ at the levels of 10^{-6}–10^{-8}. Tests sensitive to the pure gravity couplings include experiments with atom interferometers, torsion pendula, and lunar and satellite laser ranging experiments. Bounds on $\bar{s}_{\mu\nu}$ coefficients at levels of 10^{-6}–10^{-9} have been obtained. In addition to these gravity tests, highly sensitive tests attempting to detect spacetime torsion can be achieved by searching for its couplings to fermions [23.164]. Bounds on torsion components down to levels of 10^{-31} GeV have been obtained in this way.

23.4 Summary and Conclusions

Interest in the idea of Lorentz violation has steadily increased over the past two decades. This is due to theoretical advances showing that Lorentz breaking can provide unique signals of Planck-scale physics and quantum-gravity effects as well as to experimental advances that have led to new high-precision tests of LLI. The development and use of the SME as the theoretical framework describing Lorentz violation in the context of field theory has led to a comprehensive and multidisciplinary approach to testing LLI that spans most of the particle sectors in the SM.

The underlying premise of the ME is that field theory and the SM are correct descriptions of particle interactions at low energies. Therefore, any indications of Lorentz violation should show up as small corrections in the context of effective field theory. The SME is constructed as the most general effective field theory that incorporates Lorentz violation. It contains all known particle fields and gravitational interactions as well as all observer-independent terms that break LLI. As an incremental first step, the mSME and its QED limit, which maintain gauge invariance and power-counting renormalizability, were constructed in the 1990s. These have been used extensively to search for leading-order signals of Lorentz and CPT violation. More recently, a systematic approach to constructing the nonminimal sectors of SME have been worked out for certain particle species, and experimental bounds of these terms are being obtained as well [23.165, 166].

As a comprehensive theoretical framework, the SME allows for investigations of theoretical issues related to the idea of Lorentz violation. Specifically, for the case of spontaneous Lorentz breaking, investigations of the fate of the NG modes and the possibility of Higgs masses and a Higgs mechanism have been carried out. It has been shown that spontaneous Lorentz violation is accompanied by spontaneous diffeomorphism breaking, and up to 10 NG modes can appear in principle. These modes can comprise 10 of the 16 degrees of freedom of the vierbein, which in a Lorentz-invariant theory are gauge degrees of freedom. The fate of the NG modes is found to depend on the type of spacetime geometry in the underlying theory. At leading order in Minkowski and Riemann spacetimes, it is found that the NG modes can propagate like photons in a fixed axial gauge. However, in Riemann–Cartan spacetimes, the possibility exists that the spin connection can absorb the NG modes in a gravitational version of the Higgs mechanism. In addition, the potential inducing spontaneous Lorentz violation can provide mass terms for the metric excitations. These features create new possibilities for constructing models with spontaneous Lorentz violation in the context of massive gravity.

The main application of the SME has been in phenomenological investigations of Lorentz and CPT symmetry. High precision tests have been performed in most of the primary particle sectors in the SM. These include experiments in QED and atomic systems, astrophysical tests, and laboratory tests at nuclear and particle facilities. The generality of the SME allows comparisons across different types of experiments involving the same particle species. These tests have greatly improved the sensitivity to which Lorentz and CPT symmetry is known to hold, although many particle sectors, particularly those beyond leading order, remain to be probed. As a comparison of some of the bounds obtained to date at leading order, a summary of some bounds on $\tilde{b}_J$ coefficients in the minimal SME is given in Table 23.2. These bounds are within the range of sensitivity associated with suppression factors arising from the Planck scale. A more complete set of tables for the full SME has been published in the Indiana University data tables [23.88].

Table 23.2 Summary of leading-order bounds for the coefficient $\tilde{b}_J$

Experiment	Sector	Parameter ($J = X, Y$)	Bound (GeV)
Penning trap	Electron	$\tilde{b}^e_J$	10^{-24}
K-He dual maser	Neutron	$\tilde{b}^n_J$	10^{-33}
H maser	Proton	$\tilde{b}^p_J$	10^{-27}
Muonium	Muon	$\tilde{b}^\mu_J$	10^{-23}
Spin pendulum	Electron	$\tilde{b}^e_J$	10^{-31}

References

23.1 V.A. Kostelecký, R. Potting: Phys. Rev. D **51**, 3923 (1995)
23.2 D. Colladay, V.A. Kostelecký: Phys. Rev. D **55**, 6760 (1997)
23.3 D. Colladay, V.A. Kostelecký: Phys. Rev. D **58**, 116002 (1998)
23.4 V.A. Kostelecký, R. Lehnert: Phys. Rev. D **63**, 065008 (2001)
23.5 V.A. Kostelecký: Phys. Rev. D **69**, 105009 (2004)
23.6 J. Schwinger: Phys. Rev. **82**, 914 (1951)
23.7 J.S. Bell: Proc. R. Soc. A **231**, 479 (1955)
23.8 W. Pauli: Exclusion principle, lorentz group and reflection of space-time and charge. In: *Neils Bohr and the Development of Physics*, ed. by W. Pauli (McGraw-Hill, New York 1955) p. 30
23.9 G. Lüders: Ann. Phys. **2**, 1 (1957)
23.10 O.W. Greenberg: Phys. Rev. Lett. **89**, 231602 (2002)
23.11 F.W. Hehl, P. Von Der Heyde, G.D. Kerlick, J.M. Nester: Rev. Mod. Phys. **48**, 393 (1976)
23.12 I.L. Shapiro: Phys. Rep. **357**, 113 (2002)
23.13 V.A. Kostelecký, S. Samuel: Phys. Rev. D **39**, 683 (1989)
23.14 V.A. Kostelecký, S. Samuel: Phys. Rev. Lett. **63**, 224 (1989)
23.15 V.A. Kostelecký, S. Samuel: Phys. Rev. D **40**, 1886 (1989)
23.16 V.A. Kostelecký, R. Potting: Nucl. Phys. B **359**, 545 (1991)
23.17 V.A. Kostelecký, S. Samuel: Phys. Rev. Lett. **66**, 1811 (1991)
23.18 V.A. Kostelecký, R. Potting: Phys. Lett. B **381**, 89 (1996)
23.19 V.A. Kostelecký, R. Potting: Phys. Rev. D **63**, 046007 (2001)
23.20 R. Bluhm: Lect. Notes Phys. **702**, 191 (2006)
23.21 V.A. Kostelecký (Ed.): *CPT and Lorentz Symmetry IV* (World Scientific, Singapore 2008)
23.22 V.A. Kostelecký (Ed.): *CPT and Lorentz Symmetry V* (World Scientific, Singapore, 2011)
23.23 S. Liberati, D. Mattingly: Lorentz breaking effective field theory models for matter and gravity: Theory and observational constraints, arXiv:1208.1071 (2009)
23.24 O. Bertolami, D. Colladay, V.A. Kostelecky, R. Potting: Phys. Lett. B **395**, 178 (1997)
23.25 A. Connes, M. Douglas, A. Schwartz: J. High Energy Phys. **02**, 003 (1998)
23.26 I. Mocioiu, M. Pospelov, R. Roiban: Phys. Lett. B **489**, 390 (2000)
23.27 S.M. Carroll, J.A. Harvey, V.A. Kostelecky, C.D. Lane, T. Okamoto: Phys. Rev. Lett. **87**, 141601 (2001)
23.28 Z. Guralnik, R. Jackiw, S.Y. Pi, A.P. Polychronakos: Phys. Lett. B **517**, 450 (2001)
23.29 C.E. Carlson, C.D. Carone, R.F. Lebed: Phys. Lett. B **518**, 201 (2001)
23.30 A. Anisimov, T. Banks, M. Dine, M. Graesser: Phys. Rev. D **65**, 085032 (2002)
23.31 R. Gambini, J. Pullin: Phys. Rev. D **59**, 124021 (1999)
23.32 C.P. Burgess, J. Cline, E. Filotas, J. Matias, G.D. Moore: J. High Energy Phys. **0203**, 043 (2002)
23.33 J. Alfaro, H.A. Morales-Técotl, L.F. Urrutia: Phys. Rev. D **66**, 124006 (2002)
23.34 G. Amelino-Camelia: Mod. Phys. Lett. A **17**, 899 (2002)
23.35 D. Sudarsky, L. Urrutia, H. Vucetich: Phys. Rev. D **68**, 024010 (2003)
23.36 V.A. Kostelecký, R. Lehnert, M. Perry: Phys. Rev. D **68**, 123511 (2003)
23.37 A.R. Frey: J. High Energy Phys. **0304**, 012 (2003)
23.38 F.W. Stecker: Astropart. Phys. **20**, 85 (2003)
23.39 J.D. Bjorken: Phys. Rev. D **67**, 043508 (2003)
23.40 R. Myers, M. Pospelov: Phys. Rev. Lett. **90**, 211601 (2003)
23.41 J. Cline, L. Valcárcel: J. High Energy Phys. **0403**, 032 (2004)
23.42 N.E. Mavromatos: Nucl. Instrum. Methods B **214**, 1 (2004)
23.43 C.D. Froggatt, H.B. Nielsen: Ann. Phys. **14**, 115 (2005)
23.44 W.A. Christiansen, Y.J. Ng, H. van Dam: Phys. Rev. Lett. **96**, 051301 (2006)
23.45 W.A. Christiansen, Y.J. Ng, D.J.E. Floyd, E.S. Perlman: Phys. Rev. D **83**, 084003 (2011)
23.46 D. Mattingly: Phys. Rev. D **77**, 125021 (2008)
23.47 G. Amelino-Camelia, L. Freidel, J. Kowalski-Glikman, L. Smolin: Phys. Rev. D **84**, 084010 (2011)
23.48 H.P. Robertson: Rev. Mod. Phys. **21**, 378 (1949)
23.49 R. Mansouri, R.U. Sexl: Gen. Relativ. Gravit. **8**, 497 (1977)
23.50 C.N. Will: *Theory and experimentation in Gravitational Physics* (Cambridge Univ. Press, Cambridge 1993)
23.51 V.A. Kostelecký, C.D. Lane: J. Math. Phys. **40**, 6245 (1999)
23.52 D. Bao, S.-S. Chern, Z. Shen: *An Introduction to Riemann-Finsler Geometry* (Springer, New York 2000)
23.53 V.A. Kostelecký: Phys. Lett. B **701**, 470 (2011)
23.54 V.A. Kostelecký, N. Russell, R. Tso: Phys. Lett. B **716**, 470 (2012)
23.55 Y. Nambu: Phys. Rev. Lett. **4**, 380 (1960)
23.56 J. Goldstone: Nuovo. Cim. **19**, 154 (1961)
23.57 J. Goldstone, A. Salam, S. Weinberg: Phys. Rev. **127**, 965 (1962)
23.58 F. Englert, R. Brout: Phys. Rev. Lett. **13**, 321 (1964)
23.59 P.W. Higgs: Phys. Rev. Lett. **13**, 508 (1964)
23.60 G.S. Guralnik, C.R. Hagen, T.W.B. Kibble: Phys. Rev. Lett. **13**, 585 (1964)
23.61 P.A.M. Dirac: Proc. R. Soc. A **209**, 291 (1951)
23.62 Y. Nambu: Prog. Theor. Phys. Suppl. Extra **68**, 190–195 (1968)

23.63 J.D. Bjorken: Ann. Phys. **24**, 174 (1963)
23.64 R. Bluhm, V.A. Kostelecký: Phys. Rev. D **71**, 065008 (2005)
23.65 R. Bluhm, S.-H. Fung, V.A. Kostelecký: Phys. Rev. D **77**, 065020 (2008)
23.66 T. Jacobson, D. Mattingly: Phys. Rev. D **64**, 024028 (2001)
23.67 P. Kraus, E.T. Tomboulis: Phys. Rev. D **66**, 045015 (2002)
23.68 J.W. Moffat: Int. J. Mod. Phys. D **12**, 1279 (2003)
23.69 O. Bertolami, J. Paramos: Phys. Rev. D **72**, 044001 (2005)
23.70 J.L. Chkareuli, C.D. Froggatt, J.G. Jejelava, H.B. Nielsen: Nucl. Phys. B **796**, 211 (2008)
23.71 B. Altschul, V.A. Kostelecký: Phys. Lett. B **628**, 106 (2005)
23.72 R. Bluhm, N. Gagne, R. Potting, A. Vrublevskis: Phys. Rev. D **77**, 125007 (2008)
23.73 S.M. Carroll, T.R. Dulaney, M.I. Gresham, H. Tam: Phys. Rev. D **79**, 065011 (2009)
23.74 M.D. Seifert: Phys. Rev. D **79**, 124012 (2009)
23.75 M.D. Seifert: Phys. Rev. D **81**, 065010 (2010)
23.76 O.J. Franca, R. Montemajor, L.F. Urrutia: Phys. Rev. D **85**, 085008 (2012)
23.77 V.A. Kostelecký, R. Potting: Gen. Relativ. Gravit. **37**, 1675 (2005)
23.78 V.A. Kostelecký, R. Potting: Phys. Rev. D **79**, 065018 (2009)
23.79 B. Altschul, Q.G. Bailey, V.A. Kostelecký: Phys. Rev. D **81**, 065028 (2010)
23.80 S.M. Carroll, E.A. Lim: Phys. Rev. D **70**, 123525 (2004)
23.81 S. Kanno, J. Soda: Phys. Rev. D **74**, 063505 (2006)
23.82 E.A. Lim: Phys. Rev. D **71**, 063504 (2005)
23.83 P.G. Ferreira, B.M. Gripaios, R. Saffari: Phys. Rev. D **75**, 044014 (2007)
23.84 B. Li, D.F. Mota, J.D. Barrow: Phys. Rev. D **77**, 024032 (2008)
23.85 T.R. Dulaney, M.I. Gresham, M.B. Wise: Phys. Rev. D **77**, 083510 (2008)
23.86 J. Beltran Jimenez, A.L. Maroto: Phys. Rev. D **80**, 063513 (2009)
23.87 V.A. Kostelecký, C.D. Lane: Phys. Rev. D **60**, 116010 (1999)
23.88 V.A. Kostelecký, N. Russell: Rev. Mod. Phys. **83**, 11 (2011), arXiv:0801.0287
23.89 R. Bluhm, V.A. Kostelecký, N. Russell: Phys. Rev. Lett. **79**, 1432 (1997)
23.90 R. Bluhm, V.A. Kostelecký, N. Russell: Phys. Rev. D **57**, 3932 (1998)
23.91 H. Dehmelt, R. Mittleman, R.S. van Dyck Jr., P. Schwinberg: Phys. Rev. Lett. **83**, 4694 (1999)
23.92 R.K. Mittleman, I.I. Ioannou, H.G. Dehmelt, N. Russell: Phys. Rev. Lett. **83**, 2116 (1999)
23.93 G. Gabrielse, A. Khabbaz, D.S. Hall, C. Heimann, H. Kalinowsky, W. Jhe: Phys. Rev. Lett. **82**, 3198 (1999)
23.94 R. Bluhm, V.A. Kostelecký: Phys. Rev. Lett. **84**, 1381 (2000)
23.95 L.-S. Hou, W.-T. Ni, Y.-C.M. Li: Phys. Rev. Lett. **90**, 201101 (2003)
23.96 B.R. Heckel, C.E. Cramer, T.S. Cook, E.G. Adelberger, S. Schlamminger, U. Schmidt: Phys. Rev. Lett. **97**, 021603 (2006)
23.97 B.R. Heckel, E.G. Adelberger, C.E. Cramer, T.S. Cook, S. Schlamminger, U. Schmidt: Phys. Rev. D **78**, 092006 (2008)
23.98 R. Bluhm, V.A. Kostelecký, N. Russell: Phys. Rev. Lett. **82**, 2254 (1999)
23.99 D. Bear, R.E. Stoner, R.L. Walsworth, V.A. Kostelecky, C.D. Lane: Phys. Rev. Lett. **85**, 5038 (2000)
23.100 D.F. Phillips, M.A. Humphrey, E.M. Mattison, R.E. Stoner, R.F.C. Vessot, R.L. Walsworth: Phys. Rev. D **63**, 111101 (2001)
23.101 M.A. Humphrey, D.F. Phillips, E.M. Mattison, R.F.C. Vessot, R.E. Stoner, R.L. Walsworth: Phys. Rev. A **68**, 063807 (2003)
23.102 F. Cane, D. Bear, D.F. Phillips, M.S. Rosen, C.L. Smallwood, R.E. Stoner, R.L. Walsworth, V.A. Kostelecky: Phys. Rev. Lett. **93**, 230801 (2004)
23.103 B. Altschul: Phys. Rev. D **75**, 041301 (2007)
23.104 C. Gemmel, W. Heil, S. Karpuk, K. Lenz, Yu. Sobolev, K. Tullney, M. Burghoff, W. Kilian, S. Knappe-Grüneberg, W. Müller: Phys. Rev. D **82**, 111901 (2010)
23.105 B. Altschul: Phys. Rev. D **79**, 061702 (2009)
23.106 M. Smiciklas, J.M. Brown, L.W. Cheuk, M.V. Romalis: Phys. Rev. Lett. **107**, 171604 (2011)
23.107 J.M. Brown, S.J. Smullin, T.W. Kornack, M.V. Romalis: Phys. Rev. Lett. **105**, 151604 (2010)
23.108 R. Bluhm, V.A. Kostelecky, C.D. Lane, N. Russell: Phys. Rev. Lett. **88**, 090801 (2002)
23.109 R. Bluhm, V.A. Kostelecky, C.D. Lane, N. Russell: Phys. Rev. D **68**, 125008 (2003)
23.110 B. Altschul: Phys. Rev. D **81**, 041701 (2010)
23.111 ALPHA Collaboration: ALPHA Antihydrogen Experiment, CPT, Lorentz Symmetry V, ed. by V.A. Kostelecký (World Scientific, Singapore 2011)
23.112 V.A. Kostelecký, M. Mewes: Phys. Rev. Lett. **87**, 251304 (2001)
23.113 V.A. Kostelecký, M. Mewes: Phys. Rev. D **66**, 056005 (2002)
23.114 V.A. Kostelecký, M. Mewes: Phys. Rev. Lett. **97**, 140401 (2006)
23.115 V.A. Kostelecký, M. Mewes: Phys. Rev. Lett. **99**, 011601 (2007)
23.116 H. Müller, P.L. Stanwix, M.E. Tobar, E. Ivanov, P. Wolf, S. Herrmann, A. Senger, E. Kovalchuk, A. Peters: Phys. Rev. Lett. **99**, 050401 (2007)
23.117 F.R. Klinkhamer, M. Risse: Phys. Rev. D **77**, 117901 (2008)
23.118 S. Herrmann, A. Senger, K. Mohle, M. Nagel, E.V. Kovalchuk, A. Peters: Phys. Rev. D **80**, 105011 (2009)
23.119 C. Eisele, A.Y. Nevsky, S. Schiller: Phys. Rev. Lett. **103**, 090401 (2009)
23.120 M.E. Tobar, E.N. Ivanov, P.L. Stanwix, J.-M.G. le Floch, J.G. Hartnett: Phys. Rev. D **80**, 125024 (2009)

23.121 M.A. Hohensee, P.L. Stanwix, M.E. Tobar, S.R. Parker, D.F. Phillips, R.L. Walsworth: Phys. Rev. D **82**, 076001 (2010)
23.122 B. Altschul: Phys. Rev. D **80**, 091901 (2009)
23.123 J.-P. Bocquet, D. Moricciani, V. Bellini, M. Beretta, L. Casano, A. D'Angelo, R. di Salvo, A. Fantini, D. Franco, G. Gervino, F. Ghio, G. Giardina, B. Girolami, A. Giusa, V.G. Gurzadyan, A. Kashin, S. Knyazyan, A. Lapik, R. Lehnert, P. Levi Sandri, A. Lleres, F. Mammoliti, G. Mandaglio, M. Manganaro, A. Margarian, S. Mehrabyan, R. Messi, V. Nedorezov, C. Perrin, C. Randieri, D. Rebreyend, N. Rudnev, G. Russo, C. Schaerf, M.L. Sperduto, M.C. Sutera, A. Turinge, V. Vegna: Phys. Rev. Lett. **104**, 241601 (2010)
23.124 S.R. Parker, M. Mewes, P.L. Stanwix, M.E. Tobar: Phys. Rev. Lett. **106**, 180401 (2011)
23.125 S.M. Carroll, G.B. Field, R. Jackiw: Phys. Rev. D **41**, 1231 (1990)
23.126 S. Coleman, S.L. Glashow: Phys. Rev. D **59**, 116008 (1999)
23.127 S.T. Scully, F.W. Stecker: Astroparticle Phys. **31**, 220 (2009)
23.128 X.-J. Bi, Z. Cao, Y. Li, Q. Yuan: Phys. Rev. D **79**, 083015 (2009)
23.129 K. Greisen: Phys. Rev. Lett. **16**, 748 (1966)
23.130 G.T. Zatsepin, V.A. Kuz'min: Cosmic Rays (Moscow) **11**(11), 45–47 (1969)
23.131 V.A. Kostelecký: Phys. Rev. Lett. **80**, 1818 (1998)
23.132 V.A. Kostelecký: Phys. Rev. D **61**, 016002 (2000)
23.133 V.A. Kostelecký: Phys. Rev. D **64**, 076001 (2001)
23.134 J.M. Link, et al.: Phys. Lett. B **556**, 7 (2003)
23.135 BaBar Collaboration: Phys. Rev. Lett. **100**, 131802 (2008)
23.136 A. Di Domenico, K.L.O.E. collaboration: Found. Phys. **40**, 852 (2010)
23.137 V.A. Kostelecký, R.J. Van Kooten: Phys. Rev. D **82**, 101702 (2010)
23.138 R. Bluhm, V.A. Kostelecký, C.D. Lane: Phys. Rev. Lett. **84**, 1098 (2000)
23.139 V.W. Hughes, M. Grosse Perdekamp, D. Kawall, W. Liu, K. Jungmann, G. zu Putlitz: Phys. Rev. Lett. **87**, 111804 (2001)
23.140 Muon (g-2) Collaboration: Phys. Rev. Lett. **100**, 091602 (2008)
23.141 D. Colladay, V.A. Kostelecký: Phys. Lett. B **511**, 209 (2001)
23.142 M.A. Hohensee, R. Lehnert, D.F. Phillips, R.L. Walsworth: Phys. Rev. D **80**, 036010 (2009)
23.143 D0 Collaboration: Phys. Rev. Lett. **108**, 261603 (2012)
23.144 B. Charneski, M. Gomes, R.V. Maluf, A.J. da Silva: Phys. Rev. D **86**, 045003 (2012)
23.145 V.A. Kostelecký, M. Mewes: Phys. Rev. D **69**, 016005 (2004)
23.146 V.A. Kostelecký, M. Mewes: Phys. Rev. D **70**, 031902 (2004)
23.147 V.A. Kostelecký, M. Mewes: Phys. Rev. D **70**, 076002 (2004)
23.148 LSND Collaboration: Phys. Rev. D **72**, 076004 (2005)
23.149 J.S. Diaz, V.A. Kostelecký, M. Mewes: Phys. Rev. D **80**, 076007 (2009)
23.150 IceCube Collaboration: Phys. Rev. D **82**, 112003 (2010)
23.151 MINOS Collaboration: Phys. Rev. Lett. **105**, 151601 (2010)
23.152 J.S. Diaz, V.A. Kostelecký: Phys. Lett. B **700**, 25 (2011)
23.153 J.S. Diaz, V.A. Kostelecký: Phys. Rev. D **85**, 016013 (2012)
23.154 V. Barger, D. Marfatia, K. Whisnant: Phys. Lett. B **653**, 267 (2007)
23.155 T. Katori, MiniBooNE Collaboration: Test for Lorentz and CPT Violation with the MiniBooNE Low-Energy Excess, CPT and Lorentz Symmetry V, ed. by V.A. Kostelecký (World Scientific, Singapore 2011)
23.156 Q.G. Bailey, V.A. Kostelecký: Phys. Rev. D **74**, 045001 (2006)
23.157 J.B.R. Battat, J.F. Chandler, C.W. Stubbs: Phys. Rev. Lett. **99**, 241103 (2007)
23.158 K.-Y. Chung, S.-W. Chiow, S. Herrmann, S. Chu, H. Müller: Phys. Rev. D **80**, 016002 (2009)
23.159 H. Müller, S.-W. Chiow, S. Herrmann, S. Chu, K.-Y. Chung: Phys. Rev. Lett. **100**, 031101 (2008)
23.160 V.A. Kostelecký, J. Tasson: Phys. Rev. Lett. **102**, 010402 (2009)
23.161 V.A. Kostelecký, J. Tasson: Phys. Rev. D **83**, 016013 (2011)
23.162 M.A. Hohensee, S. Chu, A. Peters, H. Müller: Phys. Rev. Lett. **106**, 151102 (2011)
23.163 L. Iorio: Class. Quantum Gravity **29**, 175007 (2012)
23.164 V.A. Kostelecký, N. Russell, J. Tasson: Phys. Rev. Lett. **100**, 111102 (2008)
23.165 V.A. Kostelecký, M. Mewes: Phys. Rev. D **80**, 015020 (2009)
23.166 V.A. Kostelecký, M. Mewes: Phys. Rev. D **85**, 096005 (2012)

24. Relativity in GNSS

Neil Ashby

Global navigation satellite systems (GNSS) use accurate, stable atomic clocks in satellites and on the ground to provide world-wide position, velocity, and time to millions of users. Orbiting clocks have gravitational and motional frequency shifts that are so large that, without carefully accounting for numerous relativistic effects, the systems would not work. The basis for navigation using GNSS, founded on special and general relativity, includes relativistic principles, concepts and effects such as the constancy of the speed of light, relativity of synchronization, coordinate time, proper time, time dilation, the Sagnac effect, the weak equivalence principle, and gravitational frequency shifts. Additional small relativistic effects such as the coordinate slowing of light speed and the effects of tidal potentials from the moon and the sun may need to be accounted for in the future. Examples of new navigation systems that are being developed and deployed are the European GALILEO system and the Chinese BEIDOU system; these will greatly widen the impact of GNSS. This chapter discusses applications of relativistic concepts in GNSS.

Since the first deployment of high-performance atomic clocks in satellites in 1977, position, navigation, and timing have been revolutionized world-wide. The United States' global positioning system (GPS), the Russian global navigation satellite system (GLONASS – globalnaya navigatsionnaya sputnikovaya sistema), the European GALILEO system, and China's BEIDOU system will soon provide 100 or more satellites with synchronized clocks in precisely determined orbits. Each system consists of approximately 30 satellites, capable of transmitting messages that enable a receiver to accurately compute its position, velocity, and time anywhere near earth's surface. There are also numerous augmentation systems designed to provide improved reliability and accuracy. Examples are the US's WAAS (wide area augmentation system), which uses geosynchronous satellites to broadcast GPS-like signals over the continental United States, and Japan's QZSS system that uses satellites in highly eccentric orbits, enabling them to spend considerable time directly over an area of particular interest. These systems together are generally referred to as GNSS.

A vast infrastructure supports these systems: world-wide networks of receivers and organizations to mon-

itor and estimate the satellite orbits and clocks; ensembles of high-performance clocks on the ground to provide time references; industries to design, manufacture, and launch the satellites; and hundreds of millions of users with receivers of varying degrees of complexity and expense. The GPS infrastructure has been adequately described elsewhere [24.1].

The remarkable positioning precision achieved by GNSS is due to careful accounting for a number of systematic effects that would otherwise greatly degrade the results and eventually render the system useless. Among these effects are signal delays due to water vapor in the troposphere, free electrons in the ionosphere, and reflections of signals from surfaces near the receiver antenna. Unless relativistic concepts and effects on clocks and radio signals in the GNSS are taken into account, the systems will not work. This article discusses the fundamental principles of special and general relativity that provide the basis for positioning in the GNSS. The principle of equivalence is discussed in Sect. 24.2, where it is shown that to a first approximation, gravitational potentials due to the sun and the moon can be neglected in the GNSS. Relative motions of clocks and the rotation of the earth leads to the discussion of coordinate time and the Sagnac effect in Sect. 24.3. In Sect. 24.4 we discuss international atomic time (TAI) and universal coordinated time (UTC). Sections Sect. 24.5 through Sect. 24.9 discuss relativistic effects on ground-based clocks and orbiting clocks and how such effects are accounted for. Additional effects that are currently neglected are described in Sect. 24.10.

24.1 The Principle of Equivalence

The *weak* equivalence principle is based on the observed universality of free fall, namely that all objects fall with equal accelerations in a given gravitational field, independent of their internal structure, mass, or composition. Thus in a freely falling laboratory of sufficiently small extent, no experiment performed locally – entirely within the laboratory – can tell that the laboratory is in free fall. Although this has been tested only to a certain, very high level of precision [24.2], it means that even if there is no gravitational field due to nearby masses, then in a uniformly accelerating laboratory an induced gravitational field will appear that can in no way, by local measurements only, be distinguished from a real gravitational field.

Some have been tempted to think that clocks in satellites, which are momentarily on the side of the earth nearest the sun, are affected more by the sun than satellites on the side of earth away from the sun; one implication that has been put forward numerous times is that clocks in satellites nearer the sun suffer a greater shift in frequency toward the red than do clocks on the opposite side of the earth. By the principle of equivalence, however, this picture is erroneous.

The earth and its satellites are in free fall about the sun, moon, and other solar system bodies. Locally, the gravitational field due to external bodies causes acceleration, which in turn induces an equal but opposite fictitious gravitational field; these can be superimposed and they cancel to high precision near earth's center of mass. Let the total gravitational potential in the neighborhood of the earth be denoted by $\Phi(\boldsymbol{r})$; it will be the sum of earth's potential, $V(\boldsymbol{r})$, plus the potential due to external sources, $\phi_{\text{ext}}(\boldsymbol{r})$

$$\Phi(\boldsymbol{r}) = V(\boldsymbol{r}) + \phi_{\text{ext}}(\boldsymbol{r}) \,, \tag{24.1}$$

where $\boldsymbol{r} = \{x^1, x^2, x^3\}$ is a vector from the center of mass of the earth to the point of observation. We take the origin of spatial coordinates to be earth's center of mass. The distance $r = |\boldsymbol{r}|$ is small compared to the distance to any external source, so we may imagine a series expansion of the external potential about earth's center of mass

$$\begin{aligned}\Phi(\boldsymbol{r}) &= \phi(\boldsymbol{r}) + \phi_{\text{ext}}(0)\\ &\quad + \sum_{i=1}^{3} x^i \left.\frac{\partial \phi_{\text{ext}}}{\partial x^i}\right|_0 + \frac{1}{2}\sum_{i,j=1}^{3} x^i x^j \left.\frac{\partial^2 \phi_{\text{ext}}}{\partial x^i \partial x^j}\right|_0\\ &\quad + \ldots .\end{aligned} \tag{24.2}$$

The term $\phi_{\text{ext}}(0)$ represents a constant potential everywhere near the earth and affects all physical objects in the same way. It cannot be detected and thus can be ignored. The linear terms on the second line of (24.2) represent the strength of the gravitational field due to external sources and are canceled by the induced gravitational field due to the acceleration. This is not easy to prove from first principles but proofs can be found in the literature [24.3–5]. Evidence for this result is that the linear term would exert a huge effect on the oceans, whereas it is only the last term in (24.2) that gives rise to the ocean tides. For most purposes in the GNSS tidal

effects on clocks are small and can at first be neglected. The tidal effects will be discussed further in Sect. 24.9. We conclude that for GNSS, to a high degree of approximation the only gravitational potential of significance is that of the earth itself. Although the earth and its satellites fall freely in the gravitational fields of external sources, one can introduce coordinate axes with origin at earth's center of mass and axes pointing toward distant references in the cosmos; this defines a reference system which is locally very nearly inertial. In such a system clocks can be synchronized using constancy of the speed of light.

24.2 Navigation Principles in the GNSS

The principles of position determination and time transfer in the GNSS can be very simply stated. Let there be four synchronized atomic clocks which transmit sharply defined pulses from the positions $\boldsymbol{r}_j$ at times t_j, with $j = 1, 2, 3, 4$ an index labeling the different transmission events.

Then from the principle of the constancy of the speed of light

$$c^2(t - t_j)^2 = |\boldsymbol{r} - \boldsymbol{r}_j|^2\,, \quad j = 1, 2, 3, 4\,, \tag{24.3}$$

where the defined value of c is exactly 299 792 458 m/s. These four equations can be solved for the unknown space-time coordinates of the reception event, $\{\boldsymbol{r}, t\}$. Hence the principle of the constancy of c finds application as the fundamental concept on which navigation and timing in the GNSS is based. Obviously, it is necessary to specify carefully the reference frame in which the transmitter clocks are synchronized, so that (24.3) is valid.

Equation (24.3) is nonlinear. Typically solutions are obtained by linearizing, solving approximately, and then iterating until a solution converges. For example, if one guesses that the solution is $\boldsymbol{r} = \boldsymbol{r}_0 + \delta(\boldsymbol{r}), ct = ct_0 + \delta(ct)$, where the corrections $\delta(\boldsymbol{r})$ and $\delta(ct)$ are small, then linearizing the navigation equations gives

$$\boldsymbol{N}_j \cdot \delta(\boldsymbol{r}) - \delta(ct) = c(t_0 - t_j) - |\boldsymbol{r}_0 - \boldsymbol{r}_j|\,, \tag{24.4}$$

where $\boldsymbol{N}_j$ is a unit vector from the j-th satellite to the assumed receiver position. Four such equations can be written in matrix form and the matrix equation can be solved for the corrections; iteration of the calculation usually converges very rapidly because the distances between receiver and satellites are large compared to the distance from earth's center to the receiver.

Equation (24.4) also allows one to estimate position uncertainties arising from uncertainties in determining the propagation time intervals or from poor satellite geometry. For example, suppose a receiver is at the geometric center of a tetrahedral satellite configuration and that timing errors from the satellites are uncorrelated and are each 10 ns (1 ns $= 10^{-9}$ s); 10 ns corresponds to a position error of 3 m in each direction resulting in an estimated position which is within a sphere of radius 4.7 m. In real navigation situations such ideal tetrahedral symmetry cannot be achieved since the earth's presence forces the received signals to come from somewhat less than 2π steradians of the sky above. The position error then crucially depends on the independence of the vectors $\boldsymbol{N}_j$; if these vectors should all lie close to some plane then the position uncertainty can be many times larger. Thus, the navigation equations play an important role in design of the satellite configuration so that such errors are minimized.

Signals transmitted to users from the satellites are right circularly polarized. Usually information is transmitted by encoding the high frequency carriers with phase reversals. The timing signals in question can then be thought of as places in the transmitted wave trains where there is a particular phase reversal of the circularly polarized electromagnetic signals. At such places the electromagnetic field tensor passes through zero; these are relativistically invariant events and, therefore, provide relatively moving observers with sequences of events that they can agree on in principle.

24.3 Rotation and the Sagnac Effect

Almost all users of GNSS are at fixed locations on the rotating earth, or else are moving very slowly over earth's surface. This led to an early design decision in the GPS to broadcast the satellite ephemerides in a model earth-centered, earth-fixed reference frame (ECEF frame), in which the model earth rotates about a fixed axis with a defined rotation rate, $\omega_E = 7.292115 \times 10^{-5}\,\mathrm{rad\,s^{-1}}$. This reference frame is desig-

nated by the symbol WGS-84; the station coordinates used to define this system have been updated several times since 1984 [24.6–8]. The latest realization is termed WGS-84(G1150) and is generally assumed to be identical to the International Terrestrial Reference Frame ITRF00 [24.8]. The differences among these frames are only a few centimeters. Other GNSS systems use their own earth-fixed reference systems. The Galileo terrestrial reference frame (GTRF) is an independent realization of the International Terrestrial Reference System (ITRS) established by the Central Bureau of the International Earth Rotation Service (IERS). For discussions of relativity, the particular choice of ECEF frame is immaterial. Also, the fact that the earth truly rotates about a slightly different axis with a variable rotation rate has little consequence for relativity and will not be discussed here. We shall simply regard the ECEF frame of the appropriate GNSS system as closely related to, or determined by, the ITRF established by the International Bureau of Weights and Measures (BIPM).

It should be emphasized that the transmitted navigation messages provide the user only with a function from which the satellite position can be calculated *in the ECEF* as a function of the transmission time. Usually, the satellite transmission times t_j are unequal, so the coordinate system in which the satellite positions are specified changes orientation from one measurement to the next. Therefore, to implement (24.3), the receiver must generally perform a different rotation for each measurement made, into some common inertial frame, so that (24.3) apply. After solving the propagation delay equations, a final rotation must usually be performed into the ECEF to determine the receiver's position. This can become exceedingly complicated and confusing. A technical note [24.9] discusses these issues in considerable detail.

Although the ECEF frame is of primary interest for navigation, it is simpler to describe many physical processes (such as electromagnetic wave propagation) in an inertial reference frame. Certainly, inertial reference frames are needed to express (24.3), whereas it would lead to serious error to assert (24.3) in the ECEF frame. A *conventional inertial frame* is frequently discussed, whose origin coincides with earth's center of mass, which is in free fall with the earth in the gravitational fields of other solar system bodies, and whose z-axis coincides with the angular momentum axis of earth at the epoch J2000.0. Such a local inertial frame may be related by a transformation of coordinates to the so-called international celestial reference frame (ICRF), an inertial frame defined by the coordinates of about 500 stellar radio sources. The center of this reference frame is the barycenter of the solar system.

Let us, therefore, consider the simplest instance of a transformation from an inertial frame, in which the space-time is Minkowskian, to a rotating frame of reference. Ignoring gravitational potentials for the moment, the metric in an inertial frame in cylindrical coordinates is

$$-\mathrm{d}s^2 = -(c\,\mathrm{d}t)^2 + \mathrm{d}r^2 + r^2\mathrm{d}\phi^2 + \mathrm{d}z^2\,, \tag{24.5}$$

and the transformation to a coordinate system $\{t', r', \phi', z'\}$ rotating at the uniform angular rate ω_{E} is

$$\begin{aligned} t &= t'\,, & r &= r'\,,\\ \phi &= \phi' + \omega_{\mathrm{E}} t'\,, & z &= z'\,. \end{aligned} \tag{24.6}$$

This results in the following well-known metric (Langevin metric) in the rotating frame

$$\begin{aligned} -\mathrm{d}s^2 = &-\left(1 - \frac{\omega_{\mathrm{E}}^2 r'^2}{c^2}\right)(c\,\mathrm{d}t')^2 \\ &+ 2\omega_{\mathrm{E}} r'^2\,\mathrm{d}\phi'\,\mathrm{d}t' + (\mathrm{d}\sigma')^2\,, \end{aligned} \tag{24.7}$$

where the abbreviated expression $(\mathrm{d}\sigma')^2 = (\mathrm{d}r')^2 + (r'\,\mathrm{d}\phi')^2 + (\mathrm{d}z')^2$ for the square of the coordinate distance has been used.

The time transformation $t = t'$ in (24.6) is deceivingly simple. It means that in the rotating frame the time variable t' is really determined in the underlying inertial frame. It is an example of coordinate time. A similar concept is used in the GNSS.

Consider a process in which observers in the rotating frame attempt to use Einstein synchronization (that is, the principle of the constancy of the speed of light) to establish a network of synchronized clocks. Light travels along a null worldline, so we may set $\mathrm{d}s^2 = 0$ in (24.7). Also, it is sufficient for this discussion to keep only terms of first order in the small parameter $\omega_{\mathrm{E}} r'/c$. Then

$$(c\,\mathrm{d}t')^2 - \frac{2\omega_{\mathrm{E}} r'^2\,\mathrm{d}\phi'(c\,\mathrm{d}t')}{c} - (\mathrm{d}\sigma')^2 = 0\,, \tag{24.8}$$

and solving for $(c\,\mathrm{d}t')$,

$$c\,\mathrm{d}t' = \mathrm{d}\sigma' + \frac{\omega_{\mathrm{E}} r'^2\,\mathrm{d}\phi'}{c}\,. \tag{24.9}$$

The quantity $r'^2\mathrm{d}\phi'/2$ is just the infinitesimal area $\mathrm{d}A'_z$ in the rotating coordinate system swept out

by a vector from the rotation axis to the light pulse and projected onto a plane parallel to the equatorial plane. Thus the total time required for light to traverse some path is

$$\int_{\text{path}} dt' = \int_{\text{path}} \frac{d\sigma'}{c} + \frac{2\omega_E}{c^2} \int_{\text{path}} dA'_z \quad \text{[light]} . \tag{24.10}$$

Observers fixed on the earth, who were unaware of earth rotation, would use just $\int d\sigma'/c$ for synchronizing their clock network. Observers at rest in the underlying inertial frame would say that this leads to significant path-dependent inconsistencies, which are proportional to the projected area encompassed by the path. Consider, for example, a synchronization process which follows earth's equator eastward around the globe. For earth, $2\omega_E/c^2 = 1.6227 \times 10^{-21}$ s/m^2 and the equatorial radius is $a_1 = 6\,378\,137$ m, so the area is $\pi a_1^2 = 1.27802 \times 10^{14}$ m^2 . Thus the last term in (24.10) is

$$\frac{2\omega_E}{c^2} \int_{\text{path}} dA'_z = 207.4\,\text{ns} . \tag{24.11}$$

Traversing the equator once eastward, the last clock in the synchronization path would lag the first clock by 207.4 ns. Traversing the equator once westward, the last clock in the synchronization path would lead the first clock by 207.4 ns. From the underlying inertial frame, this can be regarded as the additional travel time required by light to catch up to the moving reference point. Simple-minded use of Einstein synchronization in the rotating frame gives only $\int d\sigma'/c$ and thus leads to a significant error.

In an inertial frame a portable clock can be used to disseminate time. The clock must be moved so slowly that changes in the moving clock's rate due to time dilation, relative to a reference clock at rest on earth's surface, are extremely small. On the other hand, observers in a rotating frame who attempt this find that the proper time elapsed on the portable clock is affected by earth's rotation rate. Factoring (24.7), the proper time increment $d\tau$ on the moving clock is given by

$$\begin{aligned}(d\tau)^2 &= \left(\frac{ds}{c}\right)^2 \\ &= dt'^2 \left[1 - \left(\frac{\omega_E r'}{c}\right)^2 - \frac{2\omega_E r'^2 d\phi'}{c^2 dt'} - \left(\frac{d\sigma'}{c dt'}\right)^2\right] .\end{aligned} \tag{24.12}$$

For a slowly moving clock $(d\sigma'/c dt')^2 << 1$, so the last term in brackets in (24.12) can be neglected. Also, keeping only first-order terms in the small quantity $\omega_E r'/c$

$$d\tau = dt' - \frac{\omega_E r'^2 d\phi'}{c^2} , \tag{24.13}$$

which leads to

$$\int_{\text{path}} dt' = \int_{\text{path}} d\tau + \frac{2\omega_e}{c^2} \int_{\text{path}} dA'_z \quad \text{[portable clock]} . \tag{24.14}$$

This should be compared with (24.10). Path-dependent discrepancies in the rotating frame are thus inescapable whether one uses light or portable clocks to disseminate time, while synchronization in the underlying inertial frame using either process is self-consistent.

Equations (24.10) and (24.14) can be reinterpreted as a means of realizing coordinate time $t' = t$ in the rotating frame, if after performing a synchronization process appropriate corrections of the form $+2\omega_E \int_{\text{path}} dA'_z/c^2$ are applied. It is remarkable how many different ways this can be viewed. For example, from the inertial frame it appears that the reference clock from which the synchronization process starts is moving, requiring light to traverse a different path than it appears to traverse in the rotating frame. The Sagnac effect can be regarded as arising from the relativity of simultaneity in a Lorentz transformation to a sequence of local inertial frames comoving with points on the rotating earth. It can also be regarded as the difference between proper times of a slowly moving portable clock and a reference clock fixed on earth's surface.

This was recognized in the early 1980s by the Consultative Committee for the Definition of the Second and the International Radio Consultative Committee, who formally adopted procedures incorporating such corrections for the comparison of time standards located far apart on earth's surface. For GNSS it means that synchronization of the entire system of ground-based and orbiting atomic clocks is performed in the local inertial frame, or *ECI* coordinate system [24.10].

Satellite clocks can be used to compare times on two earth-fixed clocks when a single satellite is in view from both locations. This is the *common-view* method of comparison of Primary standards, whose locations on earth's surface are usually known very accurately in advance from ground-based surveys. Signals from a sin-

gle GPS satellite in common view of receivers at the two locations provide enough information to determine the time difference between the two local clocks. The Sagnac effect is very important in making such comparisons, as it can amount to hundreds of nanoseconds, depending on the geometry.

24.4 Coordinate Time and TAI

For GNSS the time variable $t' = t$ becomes a coordinate time in the rotating frame of the earth, which is realized by applying appropriate corrections while performing synchronization processes. Synchronization is thus performed in the underlying inertial frame in which self-consistency can be achieved.

With this understanding, we next describe the gravitational fields near the earth due to the earth's mass itself. Assume for the moment that earth's mass distribution is static, and that there exists a locally inertial, nonrotating, freely falling coordinate system with origin at the earth's center of mass, and write an approximate solution of Einstein's field equations in isotropic coordinates

$$\begin{aligned} -\mathrm{d}s^2 = &-\left(1+\frac{2V}{c^2}\right)(c\,\mathrm{d}t)^2 \\ &+\left(1-\frac{2V}{c^2}\right) \\ &\times\left(\mathrm{d}r^2 + r^2\,\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2\right), \end{aligned} \tag{24.15}$$

where $\{r, \theta, \phi\}$ are spherical polar coordinates and where V is the Newtonian gravitational potential of the earth, given approximately by

$$V = -\frac{GM_\mathrm{E}}{r}\left[1 - J_2\left(\frac{a_1}{r}\right)^2 P_2(\cos\theta)\right]. \tag{24.16}$$

In (24.16), $GM_\mathrm{E} = 3.986004418\times10^{14}\,\mathrm{m^3\,s^{-2}}$ is the product of earth's mass times the Newtonian gravitational constant, $J_2 = 1.0826300\times10^{-3}$ is earth's quadrupole moment coefficient, and $a_1 = 6.3781370\times10^6$ is earth's equatorial radius. (WGS-84(G1150) values of these constants are used in this article [24.8].) The angle θ is the polar angle measured downward from the axis of rotational symmetry; P_2 is the Legendre polynomial of degree 2. In using (24.15), it is an adequate approximation to retain only terms of first order in the small quantity V/c^2. Higher multipole moment contributions to (24.16) have very small effect on relativity in GNSS.

One additional expression for the invariant interval is needed, the transformation of (24.16) to a rotating, ECEF coordinate system by means of transformations equivalent to (24.6). The transformations for spherical polar coordinates are

$$\begin{aligned} &t = t', \quad r = r', \\ &\theta = \theta', \quad \phi = \phi' + \omega_\mathrm{E} t'. \end{aligned} \tag{24.17}$$

Upon performing the transformations, and retaining only terms of order $1/c^2$, the scalar interval becomes

$$\begin{aligned} -\mathrm{d}s^2 = &-\left[1+\frac{2V}{c^2}-\left(\frac{\omega_\mathrm{E} r'\sin\theta'}{c}\right)^2\right](c\,\mathrm{d}t')^2 \\ &+2\omega_\mathrm{E} r'^2\sin^2\theta'\,\mathrm{d}\phi'\,\mathrm{d}t' \\ &+\left(1-\frac{2V}{c^2}\right) \\ &\times\left(\mathrm{d}r'^2 + r'^2\,\mathrm{d}\theta'^2 + r'^2\sin^2\theta'\,\mathrm{d}\phi'^2\right). \end{aligned} \tag{24.18}$$

To the order of the calculation, this result is a simple superposition of the metric, (24.15), with the corrections due to rotation expressed in (24.17). The metric tensor coefficient g'_{00} in the rotating frame is

$$\begin{aligned} g'_{00} &= -\left[1+\frac{2V}{c^2}-\left(\frac{\omega_e r'\sin\theta'}{c}\right)^2\right] \\ &\equiv -\left(1+\frac{2\Phi}{c^2}\right), \end{aligned} \tag{24.19}$$

where Φ is the effective gravitational potential in the rotating frame, which includes the static gravitational potential of the earth and a centripetal potential term.

24.4.1 The Earth's Geoid

In (24.16) and (24.17), the rate of coordinate time is determined by atomic clocks at rest at infinity. The rate of coordinate time used in GNSS, however, is closely related to international atomic time (TAI), which is a time scale computed by the (BIPM) in Paris on the basis of inputs from hundreds of primary time standards, hydrogen masers, and other clocks from all over the world. In producing this time scale, corrections are applied to re-

duce the elapsed proper times on the contributing clocks to earth's geoid, a surface of constant effective gravitational equipotential at mean sea level in the ECEF.

Universal coordinated time (UTC) is a time scale that differs from TAI by a whole number of leap seconds. These leap seconds are inserted every so often into UTC so that UTC continues to correspond to time determined by earth's rotation. Time standards organizations which contribute to TAI and UTC generally maintain their own time scales. For example, the time scale of the US Naval Observatory, based on an ensemble of hydrogen masers and Cs clocks, is denoted UTC(USNO). GPS time is steered so that, apart from the leap second differences, it stays within 100 ns of UTC(USNO). Usually this steering is so successful that the difference between GPS time and UTC(USNO) is of order 10 ns. Receiver equipment cannot tolerate leap seconds, as such sudden jumps in time would cause receivers to lose their lock on transmitted signals, and other undesirable transients would occur.

To account for the fact that reference clocks for GNSS are not at infinity, We need to consider the rates of atomic clocks at rest on the earth's geoid. These clocks move because of the earth's spin; also, they are at varying distances from the earth's center of mass since the earth is slightly oblate. In order to proceed one needs a model expression for the shape of this surface and a value for the effective gravitational potential on this surface in the rotating frame.

For this calculation, (24.18) in the ECEF is relevant. For a clock at rest on earth, (24.18) reduces to

$$-\mathrm{d}s^2 = -\left(1 + \frac{2V}{c^2} - \frac{\omega_e^2 r'^2 \sin^2\theta'}{c^2}\right)(c\mathrm{d}t')^2 . \tag{24.20}$$

with the potential V given by (24.16).

This equation determines the radius r' of the effective equipotential geoid surface as a function of polar angle θ'. The numerical value of Φ_0 at the geoid can be determined at the equator where $\theta' = \pi/2$ and $r' = a_1$. This gives

$$\begin{aligned}\frac{\Phi_0}{c^2} &= -\frac{GM_E}{a_1 c^2} - \frac{GM_E J_2}{2a_1 c^2} - \frac{\omega_E^2 a_1^2}{2c^2}\\ &= -6.95348\times10^{-10}\\ &\quad -3.764\times10^{-13} - 1.203\times10^{-12}\\ &= -6.96927\times10^{-10} .\end{aligned} \tag{24.21}$$

There are thus three distinct contributions to this effective potential: a simple $1/r$ contribution due to the earth's mass; a more complicated contribution from the quadrupole potential, and a centripetal term due to the earth's rotation. The main contribution to the gravitational potential arises from the mass of the earth, the centripetal potential correction is about 500 times smaller, and the quadrupole correction is about 2000 times smaller. These contributions have been divided by c^2 in the above equation since the time increment on an atomic clock at rest on the geoid can be easily expressed thereby. In recent resolutions of the International Astronomical Union [24.11] a *terrestrial time* scale (TT) has been defined by defining the value $\Phi_0/c^2 = 6.969290134\times10^{-10}$. Equation (24.21) agrees with this definition to within the accuracy needed for the GNSS.

From (24.18), for clocks on the geoid,

$$\mathrm{d}\tau = \frac{\mathrm{d}s}{c} = \mathrm{d}t'\left(1 + \frac{\Phi_0}{c^2}\right) . \tag{24.22}$$

Clocks at rest on the rotating geoid run slow compared to clocks at rest at infinity by about seven parts in 10^{10}. These effects sum to about 10 000 times larger than the fractional frequency stability of a high-performance cesium clock. The shape of the geoid in this model can be obtained by setting $\Phi = \Phi_0$ and solving (24.19) for r' in terms of θ'. The first few terms in a power series in the variable $x' = \sin\theta'$ can be expressed as

$$\begin{aligned}r' &= 6\,356\,742.025 + 21\,353.642x'^2 + 39.832x'^4\\ &\quad + 0.798x'^6 + 0.003x'^8\,\mathrm{m} .\end{aligned} \tag{24.23}$$

This treatment of the gravitational field of the oblate earth is limited by the simple model of the gravitational field. Actually (24.23) estimates the shape of the so-called *reference ellipsoid*, from which the actual geoid is conventionally measured.

Better models can be found in the literature of geophysics [24.12–14]. The next term in the multipole expansion of the earth's gravity field is about a thousand times smaller than the contribution from J_2; although the actual shape of the geoid can differ from (24.23) by as much as 100 m, the effects of such terms on timing in GNSS are small. Incorporating up to 20 higher zonal harmonics in a calculation Φ_0 affects the value only in the sixth significant figure.

Observers at rest on the geoid define the unit of time in terms of the proper rate of atomic clocks. In (24.22), Φ_0 is a constant. On the left-hand side of (24.22), $\mathrm{d}\tau$ is the increment of proper time elapsed on a standard

clock at rest, in terms of the elapsed coordinate time $\mathrm{d}t$. Thus the very useful result has emerged that ideal clocks at rest on the geoid of the rotating earth all beat at the same rate. This is reasonable since the earth's surface is a gravitational equipotential surface in the rotating frame. (It is true for the actual geoid, whereas here we constructed a model.) Considering clocks at two different latitudes, the one further north will be closer to the earth's center because of the flattening – it will, therefore, be more redshifted. However, it is also closer to the axis of rotation and goes more slowly, so it suffers less second-order Doppler shift. The earth's oblateness gives rise to an important quadrupole correction. This combination of effects cancels exactly on the reference surface.

Since all clocks at rest on the geoid beat at the same rate, it is advantageous to exploit this fact to redefine the rate of coordinate time. Equation (24.15) defines the rate of coordinate time in terms of the rate of standard clocks at rest at infinity. What is needed instead is to define the rate of coordinate time by standard clocks at rest on earth's geoid. Therefore, we define a new coordinate time t'' by means of a constant rate change

$$t'' = (1 + \Phi_0/c^2)t' = (1 + \Phi_0/c^2)t\,. \qquad (24.24)$$

The correction is about seven parts in 10^{10} (see (24.21)).

When this time scale change is made, the metric of (24.18) in the earth-fixed rotating frame becomes

$$\begin{aligned}-\mathrm{d}s^2 = &-\left(1+\frac{2(\Phi-\Phi_0)}{c^2}\right)(c\,\mathrm{d}t'')^2 \\ &+ 2\omega_\mathrm{E} r'^2 \sin^2\theta'\,\mathrm{d}\phi'\,\mathrm{d}t'' \\ &+\left(1-\frac{2V}{c^2}\right) \\ &\times(\mathrm{d}r'^2 + r'^2\,\mathrm{d}\theta'^2 + r'^2\sin^2\theta'\,\mathrm{d}\phi'^2)\,.\end{aligned} \qquad (24.25)$$

where only terms of order c^{-2} have been retained. Whether $\mathrm{d}t'$ or $\mathrm{d}t''$ is used in the Sagnac cross term makes no difference since the Sagnac term is very small anyway. The same time scale change in the nonrotating ECI metric, (24.15), gives

$$\begin{aligned}-\mathrm{d}s^2 = &-\left(1+\frac{2(V-\Phi_0)}{c^2}\right)(c\,\mathrm{d}t'')^2 \\ &+\left(1-\frac{2V}{c^2}\right) \\ &\times(\mathrm{d}r^2 + r^2\,\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2)\,.\end{aligned} \qquad (24.26)$$

Equations (24.25) and (24.26) imply that the proper time elapsed on clocks at rest on the geoid (where $\Phi = \Phi_0$) is identical with the coordinate time t''. This is the correct way to express the fact that ideal clocks at rest on the geoid provide all of our standard reference clocks.

24.5 The Realization of Coordinate Time

We are now able to address the real problem of clock synchronization within GNSS. In the remainder of this paper we drop the primes on t'' and just use the symbol t, with the understanding that unit of this time is referenced to one of the realizations of UTC on the rotating geoid, but with synchronization established in an underlying, locally inertial, reference frame. The metric (24.26) will henceforth be written as

$$\begin{aligned}-\mathrm{d}s^2 = &-\left(1+\frac{2(V-\Phi_0)}{c^2}\right)(c\,\mathrm{d}t)^2 \\ &+\left(1-\frac{2V}{c^2}\right) \\ &\times(\mathrm{d}r^2 + r^2\,\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2)\,.\end{aligned} \qquad (24.27)$$

The difference $(V - \Phi_0)$ that appears in the first term of (24.27) arises because in the underlying earth-centered, locally inertial (ECI) coordinate system in which the equation is expressed, the unit of time is determined by moving clocks in a spatially dependent gravitational field.

Obviously (24.27) contains within it the well-known effects of time dilation (the apparent slowing of moving clocks) and frequency shifts due to gravitation. Due to these effects, which have an impact on the net elapsed proper time on an atomic clock, the proper time elapsing on the orbiting GNSS clocks cannot simply be used to transfer time from one transmission event to another. Path-dependent effects must be accounted for.

On the other hand, according to general relativity the coordinate time variable t of (24.27) is valid in a coordinate patch large enough to cover the earth and the GNSS satellite constellations. Equation (24.27) is an approximate solution of the field equations near the earth, which include the gravitational fields due

to earth's mass distribution. In this local coordinate patch, the coordinate time is single-valued. (It is not unique, of course, because there is still gauge freedom, but (24.27) represents a fairly simple and reasonable choice of gauge.) It is natural, therefore, to propose that the coordinate time variable t of (24.27) and (24.25) be used as a basis for synchronization in the neighborhood of the earth.

To see how this works for a slowly moving atomic clock, solve (24.26) for dt as follows. First factor out $(c\,dt)^2$ from all terms on the right-hand side

$$-\,ds^2 = -\left[1+\frac{2(V-\Phi_0)}{c^2} - \left(1-\frac{2V}{c^2}\right)\frac{dr^2+r^2\,d\theta^2+r^2\sin^2\theta\,d\phi^2)}{(c\,dt)^2}\right] \times (c\,dt)^2\,. \tag{24.28}$$

Simplify by writing the velocity in the ECI coordinate system as

$$v^2 = \frac{dr^2+r^2\,d\theta^2+r^2\sin^2\theta\,d\phi^2}{dt^2}\,. \tag{24.29}$$

Only terms of order c^{-2} need be kept so the potential term modifying the velocity term can be dropped. Then upon taking a square root, the proper time increment on the moving clock is approximately

$$d\tau = \frac{ds}{c} = \left[1+\frac{(V-\Phi_0)}{c^2}-\frac{v^2}{2c^2}\right]dt\,. \tag{24.30}$$

Finally, solving for the increment of coordinate time and integrating along the path of the atomic clock,

$$\int_{\text{path}} dt = \int_{\text{path}} d\tau\left[1-\frac{(V-\Phi_0)}{c^2}+\frac{v^2}{2c^2}\right]. \tag{24.31}$$

The relativistic effect on the clock, given in (24.30), is thus corrected by (24.31).

Suppose for a moment there were no gravitational fields. Then one could picture an underlying nonrotating reference frame, a local inertial frame, unattached to the spin of the earth, but with its origin at the center of the earth. In this nonrotating frame, a fictitious set of standard clocks is introduced, available anywhere, all of them being synchronized by the Einstein synchronization procedure, and running at agreed upon rates such that synchronization is maintained. These clocks read the coordinate time t. Next one introduces the rotating earth with a set of standard clocks distributed around upon it, possibly roving around. One applies to each of the standard clocks a set of corrections based on the known positions and motions of the clocks, given by (24.31). This generates a *coordinate clock time* in the earth-fixed, rotating system. This time is such that at each instant the coordinate clock agrees with a fictitious atomic clock at rest in the local inertial frame, whose position coincides with the earth-based standard clock at that instant. Thus coordinate time is equivalent to time which would be measured by standard clocks at rest in the local inertial frame [24.15].

When the gravitational field due to the earth is considered, the picture is only a little more complicated. There still exists a coordinate time which can be found by computing a correction for gravitational redshift, given by the first correction term in (24.31).

24.6 Effects on Satellite Clocks

For atomic clocks in satellites it is most convenient to consider the motions as they would be observed in the local ECI frame. Then the Sagnac effect becomes irrelevant. (The Sagnac effect on moving ground-based receivers must still be considered.) Gravitational frequency shifts and second-order Doppler shifts must be taken into account together. The term Φ_0 in (24.30) includes the scale correction needed in order to use clocks at rest on the earth's surface as references. Earth's quadrupole contributes to Φ_0 in the term $-GM_\mathrm{E}J_2/2a_1$ in (24.21); there it contributes a fractional rate correction of -3.76×10^{-13}. This effect must be accounted for in GNSS. Also, V is the earth's gravitational potential at the satellite's position. Fortunately the earth's quadrupole potential falls off very rapidly with distance, and up until very recently its effect on satellite vehicle (SV) clock frequency was neglected. This will be discussed in a later section, for the present we only note that earth's quadrupole potential effect on orbiting GNSS clocks is only about one part in 10^{14}.

24.6.1 Satellite Orbits

Let us assume that the satellites move along Keplerian orbits. This is a good approximation for GNSS satellites, but poor if the satellites are at low altitude. This assumption yields relations with which to simplify (24.31). Since the quadrupole (and higher multipole) parts of the earth's potential are neglected, in (24.31) the potential is $V = -GM_E/r$. Then the expressions can be evaluated using what is known about the Newtonian orbital mechanics of the satellites. Denote the satellite's orbit semimajor axis by a and eccentricity by e. Then the solution of the orbital equations is as follows: [24.16] the distance r from the center of the earth to the satellite in ECI coordinates is

$$r = a(1-e^2)/(1+e\cos f)\,. \tag{24.32}$$

The angle f, called the true anomaly, is measured from perigee along the orbit to the satellite's instantaneous position. The true anomaly can be calculated in terms of another quantity E called the eccentric anomaly, according to the relationships

$$\cos f = \frac{\cos E - e}{1 - e\cos E}\,, \tag{24.33}$$

$$\sin f = \sqrt{1-e^2}\frac{\sin E}{1-e\cos E}\,. \tag{24.34}$$

Then another way to write the radial distance r is

$$r = a(1-e\cos E)\,. \tag{24.35}$$

To find the eccentric anomaly E, one must solve the transcendental equation

$$E - e\sin E = \sqrt{\frac{GM_E}{a^3}}(t-t_p)\,, \tag{24.36}$$

where t_p is the coordinate time of perigee passage.

In Newtonian mechanics, the gravitational field is a conservative field and total energy is conserved. Using the above equations for the Keplerian orbit, one can show that the total energy per unit mass of the satellite is

$$\frac{1}{2}v^2 - \frac{GM_E}{r} = -\frac{GM_E}{2a}\,. \tag{24.37}$$

Inserting (24.37) for v^2 into (24.31) results in the following expression for the elapsed coordinate time on the satellite clock

$$\Delta t = \int_{\text{path}} d\tau \times \left[1 + \frac{3GM_E}{2ac^2} + \frac{\Phi_0}{c^2} - \frac{2GM_E}{c^2}\left(\frac{1}{a} - \frac{1}{r}\right)\right]. \tag{24.38}$$

The first two constant rate correction terms in (24.38) for GPS have the values

$$\begin{aligned}\frac{3GM_E}{2ac^2} + \frac{\Phi_0}{c^2} &= +2.5046\times 10^{-10}\\ &\quad -6.9693\times 10^{-10}\\ &= -4.4647\times 10^{-10}\,.\end{aligned} \tag{24.39}$$

The negative sign in this result means that the standard clock in orbit is beating too fast, primarily because its frequency is gravitationally blueshifted. In order for the satellite clock to appear to an observer on the geoid to beat at the chosen frequency of 10.23 MHz, the satellite clocks are adjusted lower in frequency so that the proper frequency is

$$\begin{aligned}&[1 - 4.4647\times 10^{-10}]\times 10.23\,\text{MHz}\\ &= 10.22999999543\,\text{MHz}\,.\end{aligned} \tag{24.40}$$

This adjustment is accomplished on the ground before the clock is placed in orbit. Five sources of relativistic effects contribute to this frequency offset. This effect is formally incorporated into the GPS specifications [24.17] and into GLONASS [24.18] but is not mentioned in the formal GALILEO signal-in-space specifications [24.19].

For GNSS systems other than GPS, typically some choice is made concerning the nominal period required for the satellite's ground track to repeat. For GLONASS, the satellite periods are 16/17 of the GPS satellite periods, while for GALILEO, the ground track repeats after 17 orbits, which takes 10 days. For BEIDOU it appears that the satellites in medium earth orbit (MEO) will have repeating ground tracks after 13 orbits in 10 days. Table 24.1 gives the nominal semimajor axes and the fractional frequency offsets for several of the systems.

The purpose of this frequency offset is to make corrections applied by the receiver smaller, so the job of the receiver is easier. Typically navigation messages from the satellites contain three coefficients that enable

Table 24.1 Nominal values of SV clock frequency offsets

GNSS system	a (km)	$10^{10} \times \Delta f/f$
GPS	26 562.76	−4.46473
GLONASS	25 509.64	−4.36144
GALILEO	29 601.31	−4.72191
BEIDOU (MEO)	27 910.20	−4.58538
Geosynchronous	42 164.17	−5.39151

the receiver to make corrections for satellite clock errors. These coefficients are denoted by a_0, a_1, and a_2; a_0 is a time or synchronization error correction, a_1 is a frequency correction, and a_2 is a frequency drift correction. The coefficient a_2 is seldom used. Although it is quite possible to implement a system in which this *factory frequency offset* is not applied before launch, the transmitted navigation messages would have to transmit a much larger a_1 coefficient, in which the first few bits are always the same. This would be wasteful of resources and would limit the number of bits available for real variations in the actual frequency offsets.

Figure 24.1 shows a histogram of 271 values of the a_1 coefficient transmitted by the GLONASS satellites, sampled from the GLONASS broadcast ephemeris at the beginning of each year for the last 7 years. The average of this sample is very nearly zero, with an RMS variation of about 1.6×10^{-12}. In an ideal world this number would be zero. Thus for GLONASS the frequency offsets achieved are within about 4% of the desired value.

Small frequency shifts can arise from clock drift, launch vibrations, environmental changes, and other

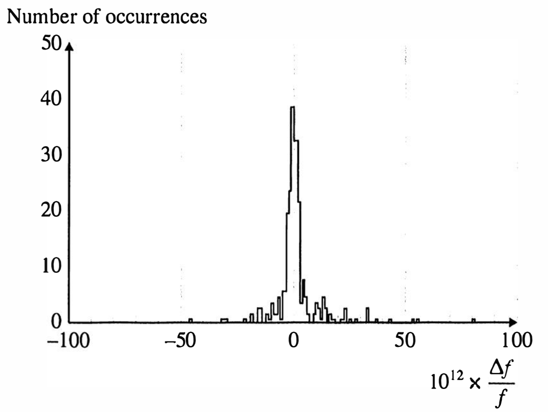


Fig. 24.1 Histogram of transmitted fractional frequency shift corrections for GLONASS. The horizontal axis is in units of 10^{-12}

unavoidable effects such as the inability to launch the satellite into an orbit with precisely the desired semi-major axis. Because of such effects, it is difficult to use GNSS clocks to measure relativistic frequency shifts.

24.6.2 The Eccentricity Correction

The last term in (24.38) may be integrated exactly by using the following expression for the rate of change of eccentric anomaly with time, which follows by differentiating (24.36)

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\sqrt{GM_\mathrm{E}/a^3}}{1 - e\cos E}\,. \tag{24.41}$$

A relativistic correction is being computed, so $\mathrm{d}s/c \simeq \mathrm{d}t$ and

$$\begin{aligned}
&\int \left[\frac{2GM_\mathrm{E}}{c^2}\left(\frac{1}{r} - \frac{1}{a}\right)\right]\frac{\mathrm{d}s}{c} \\
&\simeq \frac{2GM_\mathrm{E}}{c^2}\int (1/r - 1/a)\,\mathrm{d}t \\
&= \frac{2GM_\mathrm{E}}{ac^2}\int \mathrm{d}t\left(\frac{e\cos E}{1 - e\cos E}\right) \\
&= \frac{2\sqrt{GM_\mathrm{E}a}}{c^2}e\,(\sin E - \sin E_0) \\
&= +\frac{2\sqrt{GM_\mathrm{E}a}}{c^2}e\sin E + \text{constant}\,.
\end{aligned} \tag{24.42}$$

The constant of integration in (24.42) can be dropped since this term is lumped with other clock offset effects in the process of estimating the clock correction. The net correction for clock offset due to relativistic effects which vary in time is

$$\Delta t_\mathrm{r} = +4.4428 \times 10^{-10}\,\mathrm{s}\,\sqrt{\mathrm{m}}^{-1} e\sqrt{a}\sin E\,. \tag{24.43}$$

This correction of (24.43) is called the *eccentricity correction*; it is of the same form for all orbiting clocks and is ordinarily made by the receiver software. It represents a correction to the coordinate time as transmitted by the satellite. For a satellite of eccentricity $e = 0.01$, the maximum size of this term for GALILEO is about 24 ns. The correction is needed because of a combination of effects on the satellite clock due to gravitational frequency shift, and second-order Doppler shift, which vary due to orbit eccentricity. For the QZS-1 satellite, the amplitude of this effect is about 200 ns. Figure 24.2 gives a plot of the relativistic effect – the negative of the correction.

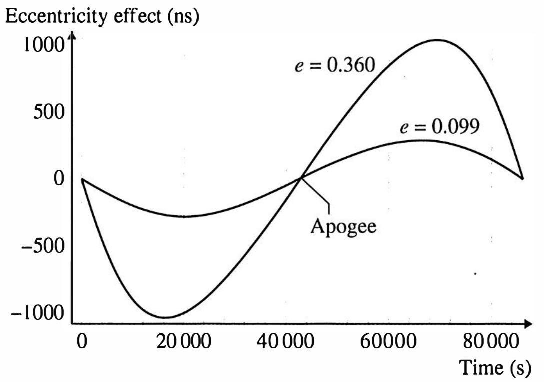


Fig. 24.2 Relativistic correction for orbital eccentricity effect, for a semimajor axis of 26 600 km

Equation (24.43) can be expressed without approximation in the following form, which is valid for Keplerian orbits,

$$\Delta t_r = +\frac{2\boldsymbol{r}\cdot\boldsymbol{v}}{c^2}\,, \qquad (24.44)$$

where $\boldsymbol{r}$ and $\boldsymbol{v}$ are the position and velocity of the satellite at the instant of transmission. This may be proved using the expressions (24.33)–(24.36) for the Keplerian orbits of the satellites. This latter form is usually used in implementations of the receiver software.

It is not necessary, in a navigation satellite system, that the eccentricity correction be applied by the receiver. It appears that the clocks in the GLONASS satellite system do have this correction applied before broadcast. In fact historically, this was dictated in the GPS by the small amount of computing power available in the early GPS satellite vehicles. It would actually make more sense to incorporate this correction into the time broadcast by the satellites; then the broadcast time events would be much closer to coordinate time – that is, GPS system time. It may now be too late to reverse this decision because of the investment that many dozens of receiver manufacturers have in their products. However, it does mean that receivers are supposed to incorporate the relativity correction; therefore if appropriate data can be obtained in raw form from a receiver one can measure this effect [24.20].

24.7 Doppler Effect

Since orbiting clocks have had their rate adjusted so they beat coordinate time, and since responsibility for correcting for the periodic relativistic effect due to eccentricity has been delegated to receivers, one must take extreme care in discussing the Doppler effect for signals transmitted from satellites. Even though second-order Doppler effects have been accounted for, for earth-fixed users there will still be a first-order (longitudinal) Doppler shift, which has to be dealt with by receivers. As is well known, in a static gravitational field coordinate frequency is conserved during propagation of an electromagnetic signal along a null geodesic. If one takes into account only the monopole and quadrupole contributions to earth's gravitational field, then the field is static and one can exploit this fact to discuss the Doppler effect.

Consider the transmission of signals from rate-adjusted transmitters orbiting on GPS satellites. Let the gravitational potential and velocity of the satellite be $V(\boldsymbol{r}_j) \equiv V_j$, and $\boldsymbol{v}_j$, respectively. Let the frequency of the satellite transmission, before the rate adjustment is done, be f_0. After taking into account the rate adjustment discussed previously, it is straightforward to show that for a receiver of velocity $\boldsymbol{v}_\mathrm{R}$ and gravitational potential V_R (in ECI coordinates), the received frequency f_R is given by

$$\begin{aligned}\frac{f_\mathrm{R}-f_0}{f_0} &= \left[\frac{-V_\mathrm{R}+v_\mathrm{R}^2/2+\Phi_0+2GM_\mathrm{E}/a+2V_j}{c^2}\right]\\ &\quad\times\frac{(1-\boldsymbol{N}\cdot\boldsymbol{v}_\mathrm{R}/c)}{(1-\boldsymbol{N}\cdot\boldsymbol{v}_j/c)}\,,\end{aligned} \qquad (24.45)$$

where $\boldsymbol{N}$ is a unit vector in the propagation direction in the local inertial frame. For a receiver fixed on the earth's rotating geoid, this reduces to

$$\frac{f_\mathrm{R}-f_0}{f_0} = \left[\frac{2GM_\mathrm{E}}{c^2}\left(\frac{1}{a}-\frac{1}{r}\right)\right]\frac{(1-\boldsymbol{N}\cdot\boldsymbol{v}_\mathrm{R}/c)}{(1-\boldsymbol{N}\cdot\boldsymbol{v}_j/c)}\,. \qquad (24.46)$$

The correction term in square brackets gives rise to the eccentricity effect. The longitudinal Doppler shift factors are not affected by these adjustments; they will be of order 10^{-5}, while the eccentricity effect is of order $e \times 10^{-10}$.

24.8 Relativity and Orbit Adjustments

To deal with satellite failures, it is common to have spares parked out of the way in orbits close to the nominal satellite orbits of the system. Performance of the clocks in these spares are monitored but not broadcast to the general user. As these spare satellites are raised or lowered in altitude to place them in assigned slots or take them out of service, their clocks suffer relativistic frequency changes from a combination of velocity changes and gravitational frequency shifts. If the initial and final orbits can be described as Keplerian orbits, (24.38) gives for the fractional frequency effect (the negative of the correction)

$$\frac{f-f_0}{f_0} = -\frac{3GM_E}{2c^2 a} - \Phi_0 \,. \tag{24.47}$$

The defined potential on the geoid, Φ_0, does not depend on satellite position. If the semimajor axis changes by a small amount δa, there will be a change in the frequency that can be adequately described by differentiating (24.47)

$$\delta\left(\frac{f-f_0}{f_0}\right) = +\frac{3GM_E}{2c^2 a^2}\delta a \,. \tag{24.48}$$

This simple equation has been very successful in predicting frequency shifts due to small changes in the semimajor axis. For a discussion of several measurements of such shifts, see [24.20]. The magnitudes of frequency shifts induced by such orbit changes are typically a few parts in 10^{13}.

The factor 3/2 in (24.48) arises from the combined effect of second-order Doppler and gravitational frequency shifts. If the semimajor axis increases, the satellite will be higher in earth's gravitational potential and will be gravitationally blueshifted more, while at the same time the satellite velocity will be reduced, reducing the size of the second-order Doppler shift (which is generally a redshift). The net effect would make a positive contribution to the fractional frequency shift.

24.9 Effects of Earth's Quadrupole Moment

Perturbations of GNSS orbits due to earth's quadrupole mass distribution are a significant fraction of the change in the semimajor axis associated with the orbit change discussed above. This raises the question whether it is sufficiently accurate to use a Keplerian orbit to describe GPS satellite orbits and estimate the semimajor axis change as though the orbit were Keplerian. In this section, we estimate the effect of earth's quadrupole moment on the orbital elements of a nominally circular orbit. Previously, such an effect on the SV clocks was neglected, and indeed it does turn out to be small. However, the effect is of the same order as the stability of the best orbiting clocks, so it is significant.

To see how large such quadrupole effects may be, we use exact calculations available in the literature, for the perturbations of the Keplerian orbital elements [24.16]. For the semimajor axis, if the eccentricity is very small the dominant contribution has a period twice the orbital period and has amplitude $3J_2a_1^2\sin^2 i/(2a)$, where a_1 is earth's equatorial radius and i is the inclination of the satellite orbit. The amplitude can be more than a kilometer.

The oscillation in the semimajor axis would significantly affect calculations of the radius at any particular time. This suggests that (24.37) needs to be reexamined in light of the periodic perturbations on the semimajor axis. Therefore, in this section we develop an approximate description of a satellite orbit, for small eccentricity, taking into account earth's quadrupole moment to first order. Terms of order $J_2 \times e$ will be neglected. This problem is nontrivial because the perturbations themselves (see, for example, the equations for mean anomaly and altitude of perigee) have factors $1/e$, which blow up as the eccentricity approaches zero. This problem is a mathematical one, not a physical one. It simply means that the observable quantities – such as coordinates and velocities – need to be calculated in such a way that finite values are obtained.

24.9.1 Conservation of Energy

The gravitational potential of a satellite at position (x, y, z) in equatorial ECI coordinates in the model under consideration here is

$$V(x,y,z) = -\frac{GM_E}{r}\left(1 - \frac{J_2a_1^2}{r^2}\left[\frac{3z^2}{2r^2} - \frac{1}{2}\right]\right) . \tag{24.49}$$

Since the force is conservative in this model (solar radiation pressure, thrust, etc., are not considered), the kinetic plus potential energy is conserved. Let ϵ be the energy per unit mass of an orbiting mass point. Then

$$\begin{aligned}\epsilon &= \text{constant}\\ &= \frac{v^2}{2} + V(x,y,z)\\ &= \frac{v^2}{2} - \frac{GM_E}{r} + V'(x,y,z)\,,\end{aligned} \tag{24.50}$$

where $V'(x,y,z)$ is the perturbing potential due to the earth's quadrupole potential.

It is shown in textbooks [24.16] that, with the help of Lagrange's planetary perturbation theory, the conservation of energy condition can be put in the form

$$\epsilon = -\frac{GM_E}{2a} + V'(x,y,z)\,, \tag{24.51}$$

where a is the perturbed (osculating) semimajor axis. In other words, for the perturbed orbit,

$$\frac{v^2}{2} - \frac{GM_E}{r} = -\frac{GM_E}{2a}\,. \tag{24.52}$$

On the other hand, the net fractional frequency shift relative to a clock at rest at infinity is determined by the second-order Doppler shift (a redshift) and a gravitational redshift. The total relativistic fractional frequency shift (relative to a reference at infinity) is

$$\frac{\Delta f}{f} = -\frac{v^2}{2} - \frac{GM_E}{r} + V'(x,y,z)\,. \tag{24.53}$$

The conservation of energy condition can be used to express the second-order Doppler shift in terms of the potential. Therefore, from perturbation theory we need expressions for the square of the velocity, for the radius r, and for the perturbing potential. We now proceed to derive these expressions. We refer to the literature [24.16] for the perturbed osculating elements. These are exactly known, to all orders in the eccentricity, and to first order in J_2. We shall need only the leading terms in eccentricity e for each element.

24.9.2 Perturbed Semimajor Axis

From [24.16], the perturbed semimajor axis in the limit of negligible eccentricity is

$$a = a_m + \frac{3J_2a_1^2}{2a_m}\sin^2 i\cos(2nt+2\omega)\,, \tag{24.54}$$

where $n=\sqrt{GM_E/a_m^3}$ is the unperturbed mean motion, a_m is the mean semimajor axis, i the mean inclination, $n=\sqrt{GM_E/a_m^3}$ the unperturbed mean motion, and ω the mean altitude of perigee.

24.9.3 Perturbed Radius

The orbit radius depends on the combination $e\cos E$ where E is the eccentric anomaly. The eccentric anomaly depends on the mean anomaly; perturbation equations for the mean anomaly have terms with a factor e^{-1}, so one must take extra care in computing the product $e\cos E$ in order to obtain a meaningful result in the limit of small eccentricity. For the perturbed radius we then obtain

$$\begin{aligned} r &= a_m(1-e_m\cos E_m)\\ &\quad - \frac{3J_2a_1^2}{2a_m}\sin^2 i\cos(2nt+2\omega)\,.\end{aligned} \tag{24.55}$$

24.9.4 Perturbed Velocity

Then conservation of energy, (24.50) gives the following expression for the velocity

$$\begin{aligned}\frac{v^2}{2} &= \frac{GM_E(1+e_m\cos E_m)}{2a_m(1-e_m\cos E_m)}\\ &\quad+\frac{3GM_EJ_2a_1^2}{2a_m^3}\left(1-\frac{3}{2}\sin^2 i\right)\\ &\quad+\frac{GM_EJ_2a_1^2}{2a_m^2}\sin^2 i\cos(2nt+2\omega)\,.\end{aligned} \tag{24.56}$$

24.9.5 Evaluation of the Perturbing Potential

Since the perturbing potential contains the small factor J_2, to leading order, we may substitute unperturbed values for r and z in $V'(x,y,z)$ which yields the expression

$$\begin{aligned}V'(x,y,z) &= -\frac{GM_EJ_2a_1^2}{2a_m^3}\left(1-\frac{3}{2}\sin^2 i\right)\\ &\quad-\frac{3GM_EJ_2a_1^2\sin^2 i}{4a_m^3}\cos(2nt+2\omega)\,.\end{aligned} \tag{24.57}$$

24.9.6 Fractional Frequency Shift

The fractional frequency shift calculation is very similar to the calculation of the energy, except that the second-

order Doppler term contributes with a *negative* sign. The result is

$$\begin{aligned}\frac{\Delta f}{f} &= -\frac{v^2}{2c^2} - \frac{GM_E}{c^2 r} + \frac{V'}{c^2} \\ &= -\frac{3GM_E}{2a_m c^2} - \frac{2GM_E}{c^2 a_m}\frac{e_m \cos E_m}{1 - e_m \cos E_m} \\ &\quad - \frac{7GM_e J_2 a_1^2}{2a_m^3 c^2}\left(1 - \frac{3}{2}\sin^2 i\right) \\ &\quad - \frac{GM_E J_2 a_1^2 \sin^2 i}{a_m^3 c^2}\cos(2nt + 2\omega)\,.\end{aligned} \tag{24.58}$$

The first term, when combined with the reference potential at earth's geoid gives rise to the*factory frequency offset*. The second term gives rise to the eccentricity effect. The third term can often be neglected. The angle of inclination for which the third term vanishes exactly is $i = 55°$. For good coverage in the temperate zones, the orbits of most satellite navigation systems have inclinations very close to this value. For GPS the last term has an amplitude

$$\frac{GM_E J_2 a_1^2 \sin^2 i}{a_m^3 c^2} = 6.95 \times 10^{-15}\,. \tag{24.59}$$

The best clocks in orbit in the GPS have stabilities of around 5 parts in 10^{15} at 1 day; this is only slightly less than the quadrupole effect, suggesting that this deterministic effect should be included in the systematic error budget.

The last periodic term in (24.58) is of a form similar to that which gives rise to the eccentricity correction, which is applied by GNSS receivers. Considering only the last periodic term, the additional time elapsed on the orbiting clock will be given by

$$\delta t_{J_2} = \int_{\text{path}} \mathrm{d}t \left[-\frac{GM_E J_2 a_1^2 \sin^2 i}{a_m^3 c^2} \times \cos(2nt + 2\omega)\right]. \tag{24.60}$$

Upon integrating and dropping the constant of integration (assuming as usual that such constant time offsets are lumped with other contributions) gives the periodic relativistic effect on the elapsed time of the SV clock due to earth's quadrupole moment

$$\delta t_{J_2} = -\sqrt{\frac{GM_E}{a_m^3}}\frac{J_2 a_1^2 \sin^2 i}{2c^2} \times \sin(2nt + 2\omega)\,. \tag{24.61}$$

The correction which should be applied by the receiver is the *negative* of this expression

$$\delta t_{J_2}(\text{correction}) = \sqrt{\frac{GM_E}{a_m^3}}\frac{J_2 a_1^2 \sin^2 i}{2c^2} \times \sin(2nt + 2\omega)\,. \tag{24.62}$$

The phase of this correction is zero when the satellite passes through earth's equatorial plane going northwards.

24.9.7 Effect of Other Solar System Bodies

One set of effects that has been rediscovered many times are the redshifts due to other solar system bodies. The principle of equivalence implies that sufficiently near the earth, there can be no linear terms in the effective gravitational potential due to other solar system bodies, because the earth and its satellites are in free fall in the fields of all these other bodies. The net effect locally can only come from tidal potentials, the third terms in the Taylor expansions of such potentials about the origin of the local freely falling frame of reference. Such tidal potentials from the sun, at a distance r from earth, are of order $GM_\odot r^2/R^3$, where R is the earth-sun distance [24.3]. The gravitational frequency shift of most GNSS satellite clocks from such potentials is a few parts in 10^{16}. However, this potential causes orbit perturbations of GNSS satellites that change both the radius in the main potential term $-GM_\odot/r$ and in the velocity; thus there are three contributions to the net frequency shift arising from this tidal potential. The geometry is complicated because earth's equatorial plane, the satellite orbital plane, and the ecliptic are inclined with respect to each other. Furthermore, there is a similar set of contributions from the moon's tidal potential that is larger and that can add to or subtract from solar tidal effects in a time-dependent manner. The net fractional frequency shift on a GALILEO satellite is estimated to be about five parts in 10^{15}.

24.10 Secondary Relativistic Effects

There are several additional significant relativistic effects which must be considered at the level of accuracy of a few centimeters (which corresponds to 100 ps of delay). Many investigators are modeling systematic effects down to the millimeter level, so these effects, which are currently not sufficiently large to affect navigation, may have to be considered in the future.

24.10.1 Signal Propagation Delay

The Shapiro signal propagation delay may be easily derived in the standard way from the metric, (24.25), which incorporates the choice of coordinate time rate expressed by the presence of the term in Φ_0/c^2. Setting $ds^2 = 0$ and solving for the increment of coordinate time along the path increment

$$d\sigma = \sqrt{dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta \, d\phi^2}$$

gives

$$dt = \frac{1}{c}\left[1 - \frac{2V}{c^2} + \frac{\Phi_0}{c^2}\right] d\sigma \,. \quad (24.63)$$

The time delay is sufficiently small that quadrupole contributions to the potential (and to Φ_0) can be neglected. Integrating along the straight line path a distance l between the transmitter and receiver gives for the time delay

$$\Delta t_{\text{delay}} = \frac{\Phi_0}{c^2}\frac{l}{c} + \frac{2GM_E}{c^3}\ln\left[\frac{r_1 + r_2 + l}{r_1 + r_2 - l}\right], \quad (24.64)$$

where r_1 and r_2 are the distances of transmitter and receiver from earth's center. The second term is the usual expression for the Shapiro time delay. It is modified for GNSS by a term of opposite sign (Φ_0 is negative), due to the choice of coordinate time rate, which tends to cancel the logarithm term. The net effect for a satellite to earth link is less than 2 cm and for most purposes can be neglected. One must keep in mind, however, that in the main term, l/c, l is a coordinate distance and further small relativistic corrections are required to convert it to a proper distance.

24.10.2 Effect on Geodetic Distance

At the level of a few millimeters, spatial curvature effects should be considered. For example, using the metric (24.26), the proper distance between a point at radius r_1 and another point at radius r_2 directly above the first is approximately

$$\int_{r_1}^{r_2} dr\left[1 + \frac{GM_E}{c^2 r}\right] = r_2 - r_1 + \frac{GM_E}{c^2}\ln\left(\frac{r_2}{r_1}\right). \quad (24.65)$$

Between earth's surface and the radius of a geosynchronous satellite, the difference between proper distance and coordinate distance, and between the earth's surface and the radius of GPS satellites, is approximately 8 mm. Effects of this order of magnitude would enter, for example, in the comparison of laser ranging to GPS satellites, with numerical calculations of satellite orbits based on relativistic equations of motion using coordinate times and coordinate distances.

24.10.3 Phase Wrap-Up

Transmitted signals from GNSS satellites are right circularly polarized and thus have negative helicity. For a receiver at a fixed location, the electric field vector rotates counterclockwise, when observed facing into the arriving signal. Let the angular frequency of the signal be ω in an inertial frame, and suppose the receiver spins rapidly with angular frequency Ω, which is parallel to the propagation direction of the signal. The antenna and signal electric field vector rotate in opposite directions and thus the received frequency will be $\omega + \Omega$. In the literature this is described in terms of an accumulation of phase called *phase wrap-up*. This effect has been experimentally measured with receivers spinning at rotational rates as low as 8 Hz [24.21, 22]. It is similar to an additional Doppler effect; it does not affect navigation if four signals are received simultaneously by the receiver as in (24.1).

24.11 Conclusions

GNSS is a remarkable laboratory for applications of the concepts of special and general relativity. It is also valuable as an outstanding source of pedagogical examples. It is particularly important to confirm that the basis for synchronization is on a firm conceptual foundation.

Plans are being made to put a laser-cooled clock having stability of $5 \times 10^{-14}/\sqrt{\tau}$ and accuracy of 1×10^{16}, on the international space station [24.23]. This will open up additional possibilities for testing relativity as well as for making improvements in GNSS.

References

24.1 J.J. Spilker Jr., B.W. Parkinson: Overview of GPS operation and design. In: *Global Positioning System: Theory and Applications*, (American Institute of Aeronautics and Astronautics, Washington 1996) p. 33

24.2 S. Schlamminger, K.-Y. Choi, T.A. Wagner, J.H. Gundlach, E.G. Adelberger: Test of the equivalence principle using a rotating torsion balance, Phys. Rev. Lett. **100**, 041101 (2008)

24.3 N. Ashby, B. Bertotti: Relativistic effects in local inertial frames, Phys. Rev. D **34**, 2246–2258 (1986)

24.4 R.A. Nelson: J., Math Phys. **28**, 2379–2383 (1987)

24.5 R.A. Nelson: J., Math Phys. **35**, 6224–6225 (1994)

24.6 S. Malys, J. Slater: Maintenance and enhancement of the world geodetic system 1984, Proc. ION-GPS-94 (1994) pp. 17–24

24.7 National Imagery and Mapping Agency Technical Report 8350.2, World Geodetic System 1984, Third Edition, Amendment 1, NIMA Stock No. DMATR83502WGS84, NSN 7643-01-402-0347

24.8 Addendum to NIMA TR 8350.2: Implementation of the World Geodetic System 1984 (WGS 84), Reference Frame G1150 (National Geospatial-Intelligence Agency 2001)

24.9 N. Ashby, M. Wess: *Global Positioning Receivers and Relativity*, NIST Technical Note, Vol. 1385 (U.S. Government Printing Office, Washington 1999)

24.10 N. Ashby: *An Earth-Based Coordinate Clock Network*, NBS Technical Note, Vol. 659 (U.S. Dept. of Commerce, U. S. Government Printing Office, Washington 1975), 20402 (S. D. Catalog # C13:46:659)

24.11 G. Kaplan: The IAU Resolutions on astronomical reference systems, time scales, and earth rotation models, US Naval Observatory Circular **179** (2005)

24.12 K. Lambeck: *Geophysical Geodesy*, Oxford Science Publications (Clarendon, Oxford 1988) pp. 13–18

24.13 G.D. Garland: *The Earth's Shape and Gravity* (Pergamon, New York 1965)

24.14 N. Ashby, J.J. Spilker Jr.: Introduction to relativistic effects on the Global Positioning System. In: *Global Positioning System: Theory and Applications*, (American Institute of Aeronautics and Astronautics, Washington 1996) pp. 623–697

24.15 N. Ashby, D.W. Allan: Practical implications of relativity for a global coordinate time scale, Radio Sci. **14**, 649–669 (1979)

24.16 P. Fitzpatrick: *The Principles of Celestial Mechanics* (Academic, New York 1970)

24.17 NAVSTAR GPS Space Segment/Navigation User Interfaces, ICD-GPS-200, Revision C (ARINC Research Corporation, Fountain Valley 1993)

24.18 GLONASS Interface Control Document, Edition 5.1 (Moscow 2008) p. 14

24.19 GALILEO Open Service Signal in Space Interface Control Document, Issue 1.1 (2010)

24.20 Ashby N.: Relativistic Effects in the global positioning system, available online at http://www.relativitylivingreviews.org/Articles/lrr-2003-1 (2007)

24.21 J.D. Kraus: *Antennas*, 2nd edn. (McGraw-Hill, New York 1988), reprinted by Cygnus-Quasar Books, Powell, Ohio

24.22 A.K. Tetewsky, F.E. Mullen: Effects of platform rotation on GPS with implications for GPS simulators, Proc. ION-GPS-96 (1996) pp. 1917–1925

24.23 C. Salomon, L. Cacciapuoti, N. Dimarcq: atomic clock ensemble in space: Fundamental physics and applications, Int. Jour. Mod. Phys. D **16**, 2511 (2007)

25. Quasi-local Black Hole Horizons

Badri Krishnan

This chapter introduces the subject of quasi-local horizons at a level suitable for graduate students who have taken a first course on general relativity. It reviews properties of trapped surfaces and trapped regions in some simple examples, followed by general properties of trapped surfaces including their stability properties. This is followed by a discussion of dynamical-, trapping-, and isolated-horizons with some illustrative applications.

Part D | 25.1

25.1 Overview

The first conception of a black hole was due to Michell and Laplace in the 18th century. They viewed it as a star whose gravitational field is so strong that the Newtonian escape velocity $\sqrt{2GM/R}$ (with M and R being the mass and radius of the star, respectively) is larger than the speed of light. The condition on the escape velocity leads to the inequality $R \leq 2GM/c^2$, which, remarkably, also holds in general relativity. While such a star would have the property that not even light can escape from it, this is however a nonrelativistic concept. The speed of light is not privileged in prerelativistic physics, and a moving observer would not necessarily see it as a dark object. A more complete account of this history is given by *Hawking* and *Ellis* [25.1] including reprints of the original articles.

The history of black holes proper dates back to just after the discovery of general relativity. The first nontrivial exact solution to the Einstein equations discovered by Schwarzschild in 1916 and named after him was, in fact, a black hole. It was however more than four decades before its properties were fully appreciated. The Kruskal–Szekeres extension of the Schwarzschild solution was discovered only in 1960. This was shortly followed by the discovery of the Kerr solution in 1963 representing spinning black holes. Its global properties were explained by Carter in 1966. The Kerr–Newman solution representing charged, spinning black holes was discovered in 1965. It was in 1964 that the phrase *black hole* was first coined by John Wheeler. During the same time, there were seminal developments

in understanding the general properties of black holes beyond specific examples. This includes the study of the global properties of black hole spacetimes, the definition of event horizons, and crucially for the developments to be discussed here, the singularity theorems of Penrose and Hawking and the introduction of trapped surfaces by Penrose. This was soon followed by the understanding of black hole thermodynamics by Bekenstein, Bardeen, Carter, and Hawking in 1973, and the discovery of Hawking radiation in 1974. The cosmic censorship hypothesis was formulated by Penrose in 1979. The question of whether this is valid, i. e., if every singularity that results from the future evolution of generic, regular initial conditions is hidden behind an event horizon, is not settled and is one of the key unsolved questions in classical general relativity. The black hole uniqueness theorems which showed that the Kerr–Newman solutions are the unique globally stationary black hole solutions in Einstein–Maxwell theory in four dimensions was established in the 1980s following the work of Israel, Carter, and Robinson.

More recently, black holes have been the subject of intense study in quantum gravity where the calculation of black hole entropy has been seen as a key milestone for string theory and loop quantum gravity. There have also been important developments on the classical side where the long-standing problem of calculating the gravitational wave signal from the the merger of two black holes was finally solved numerically in 2005 by Pretorius. In an astrophysical context, black holes are believed to be engines for some of the most violent events in out universe, such as active galactic nuclei. Astronomers have succeeded in locating a large number of black hole candidates with masses ranging from a few to billions of solar masses, and the direct detection of gravitational waves from binary black hole systems is expected later this decade.

Most of these seminal developments have relied on event horizons to characterize the boundary of the black hole region (the singularity theorems are a notable exception). This is completely reasonable when we are dealing with stationary situations, but can lead us astray in dynamical situations. As we shall elaborate later, event horizons are global notions and it is in principle not possible for a mortal observer to locate them. One possible alternative is to use the notion of trapped surfaces introduced by Penrose. While not entirely local since they are closed spacelike surfaces, these provide a quasi-local alternative which an observer could, in principle, locate in order to detect the presence of a black hole. Trapped surfaces lead logically to various kinds of quasi-local horizons including isolated, dynamical, and trapping horizons. The goal of this chapter is to motivate and explain the quasi-local approach to studying black hole horizons, and to review some recent results. Somewhat surprisingly, we shall see there is still a major gap in our understanding of classical black holes in dynamical spacetimes. If we accept black holes as bonafide astrophysical objects, we still do not have a satisfactory notion of what the surface of a black hole is. Event horizons are not satisfactory because of their global properties, but there is as yet no established quasi-local alternative.

The style of this chapter will typically be to start informally with simple examples and to use them as guidance for developing general concepts and definitions. In Sect. 25.2, we shall start with the simplest black hole, i. e., the spherically symmetric Schwarzschild spacetime, and understand the properties of its black hole region. This naturally motivates the fundamental notions of event horizons and trapped surfaces, and to the boundary of the trapped region. As we shall see, the different reasonable definitions of the black hole horizon agree in Schwarzschild. This will not be the case in more general situations. Perhaps the simplest example is the imploding Vaidya spacetime which shall be our second example in Sect. 25.2.2. We shall see that at least in this simple spherically symmetric example, the location of the trapped surfaces can be determined. These two examples are then followed by general definitions of event horizons, and trapped surfaces in Sect. 25.3 which formalizes many notions introduced in Sect. 25.2. Some properties of trapped surfaces under deformations and time evolution are then discussed in Sect. 25.3.3, and this leads naturally to the notions of marginally trapped tubes, and trapping and dynamical horizons. We then restrict our attention to isolated horizons which describes the equilibrium case, when no matter or radiation is falling into the black hole (but the rest of spacetime is allowed to be dynamical). This is now well understood and we review the general formalism in Sect. 25.4 with the Kerr black hole as the prototypical example. In particular, we discuss two applications: black hole thermodynamics and the spacetime in the neighborhood of an isolated black hole. Section 25.5 reviews some results and applications for dynamical horizons, and finally Sect. 25.6 provides a summary and some open issues.

The discussion of this chapter will be mostly self-contained; though digressions into the relevant mathematical concepts will be necessarily brief. Useful references for black holes and general relativity are [25.2, 3],

and for more mathematically inclined readers [25.4] is recommended as a useful introduction to the relevant concepts in differential geometry. The discussion is meant to be accessible to physics graduate students who have taken a first course in general relativity (covering, say, the first part of the textbook by *Wald* [25.3]). In the same spirit, the list of references is not meant to be exhaustive in any sense, and is mostly biased toward reviews and pedagogical material. The selection of topics is not meant to be exhaustive; this contribution is *not* to be viewed as a broad review article. It is rather a combination of pedagogically useful examples, and a brief description of some recent results. This material will hopefully what the reader's appetite and motivate him/her to delve further into the subject.

Some words on notation are in order. A *spacetime* is a smooth four-dimensional manifold $\mathcal{M}$ with a Lorentzian metric g_{ab} with signature $(-+++)$. We shall use a combination of index-free notation and Penrose's abstract index notation for tensors [25.5]; lower-case Latin letters $a, b, c, \dots$ will denote spacetime indices. Symmetrization of indices will be denoted by round brackets, e.g., $X_{(ab)} := (X_{ab} + X_{ba})/2$, and antisymmetrization by square brackets: $X_{[ab]} := (X_{ab} - X_{ba})/2$. The derivative-operator compatible with g_{ab} will be denoted by ∇_a, and the Riemann tensor R_{abcd} will be defined by $2\nabla_{[a}\nabla_{b]}X_c = R_{abc}{}^d X_d$ for an arbitrary 1-form X_a. The Ricci tensor and scalar are, respectively, $R_{ab} = R_{acb}{}^c$ and $R = g^{ab}R_{ab}$. The coordinate derivative operator will be denoted by ∂. The exterior derivative will be either denoted by indices, such as $\nabla_{[a}X_{b]}$, or in index-free notation as $\mathrm{d}X$. The Lie derivative of an arbitrary tensor field T along a vector field X will be denoted by $\mathcal{L}_X T$. Where no confusion is likely to arise, we shall often not explicitly include the indices in geometric quantities. Unless otherwise mentioned, we shall work in geometrical units with $G = 1$ and $c = 1$. We shall often deal with submanifolds of $\mathcal{M}$; a submanifold of unit codimension will be called a hyper-surface while lower dimensional manifolds (typically these will be 2-spheres topologically) will be called surfaces. All submanifolds shall be assumed to be sufficiently smooth. Unless mentioned otherwise, we shall be working with standard general relativity in four spacetime dimensions.

25.2 Simple Examples

25.2.1 The Trapped Region in Schwarzschild Spacetime

We shall start by studying the gravitational field in the vicinity of a time-independent massive spherically symmetric body. In this section, we recall some basic properties of the Schwarzschild solution.

The Schwarzschild metric is a static, spherically symmetric solution of the vacuum Einstein equations $R_{ab} = 0$. It is usually presented as

$$\mathrm{d}s^2 = -\left(1 - \frac{2M}{r}\right)\mathrm{d}t^2 + \left(1 - \frac{2M}{r}\right)^{-1}\mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\,. \tag{25.1}$$

Here r is a radial coordinate such that the area of spheres at fixed r and t is $4\pi r^2$; these spheres can be obtained invariantly by applying rotational isometries to a given initial point in the manifold. Each of these spheres is isometric to the standard round spheres in Euclidean space, and (θ, ϕ) are the usual polar coordinates. The time coordinate is t, and the metric is explicitly time independent in these coordinates. The parameter M is the mass, and in nongeometrical units, we would have the replacement $M \to GM/c^2$. This metric turns out to be an excellent approximation to, say, the gravitational field in our solar system with the sun treated as a point mass and with $\phi(r) = GM/c^2 r$ being its Newtonian gravitational potential. The quantity $R_s := 2GM/c^2$ is known as the Schwarzschild radius, and for the sun $R_s \approx 3$ km (which agrees with the Michell–Laplace idea mentioned at the very beginning of this chapter).

The metric as written above is regular and nondegenerate for $2M < r < \infty$ and $-\infty < t < \infty$. The singularity at $r = 2M$ is not a true physical singularity [25.3] and can be removed by the transformation $(t, r) \to (v, r)$ where

$$\mathrm{d}v = \mathrm{d}t + \left(1 - \frac{2M}{r}\right)\mathrm{d}r\,. \tag{25.2}$$

In these coordinates (the ingoing Eddington–Finkelstein coordinates) the metric becomes

$$\mathrm{d}s^2 = -\left(1 - \frac{2M}{r}\right)\mathrm{d}v^2 + 2\,\mathrm{d}v\,\mathrm{d}r + r^2\,\mathrm{d}\Omega^2\,, \tag{25.3}$$

where $\mathrm{d}\Omega^2 := \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$. The metric is now regular for $r > 0$ and $-\infty < v < \infty$. We could extend the solution further by going to double null coordinates, but this shall suffice for now.

Consider now the vector fields

$$\ell = \frac{\partial}{\partial v} + \frac{1}{2}\left(1 - \frac{2M}{r}\right)\frac{\partial}{\partial r}, \quad n = -\frac{\partial}{\partial r}. \tag{25.4}$$

It is easy to check that these (future directed) vector fields are both null, i. e., $\ell\cdot\ell = n\cdot n = 0$, and $\ell\cdot n = -1$. By convention, we take ℓ to be outward pointing and n to be inward pointing; we have designated $r \to \infty$ to be *outward*.

Let us pause to recall the notions of *expansion* and *shear* of a vector field. For a timelike vector field ξ, the set of vectors orthogonal to ξ form a three-dimensional plane at each point. If A is an infinitesimal area element in this plane and λ the affine parameter along ξ, then the expansion of ξ is defined as

$$\Theta_{(\xi)} = \frac{1}{A}\frac{\mathrm{d}A}{\mathrm{d}\lambda}. \tag{25.5}$$

Since a null vector field is orthogonal to itself, it is easy to show that any vector field V satisfying $V\cdot\xi = 0$ can be written as a linear combination $V = \alpha\xi + \beta\mathbf{e}_{(1)} + \gamma\mathbf{e}_{(2)}$ where $\mathbf{e}_{(1)}$ and $\mathbf{e}_{(2)}$ are mutually orthogonal unit spacelike vectors orthogonal to ξ, and α, β, γ are real numbers. The expansion of ξ is then defined as in (25.5) above except that the relevant area element A is in the two-dimensional plane spanned by $\mathbf{e}_{(1)}$ and $\mathbf{e}_{(2)}$. An alternate expression for the expansion which is usually more useful is

$$\Theta_{(\xi)} = q^{ab}\nabla_a\xi_b, \tag{25.6}$$

where q^{ab} is the (inverse of) the Riemannian metric in the $(e_{(1)}, e_{(2)})$ plane. The trace of $\nabla_a\xi_b$ after projection is the expansion. The symmetric trace-free part and the antisymmetric parts give the shear σ_{ab} and twist ω_{ab}, respectively

$$\sigma_{ab} = \left(\nabla_{(a}\xi_{b)}\right)^{\top} - \tfrac{1}{2}\Theta_{(\xi)}q_{ab}, \quad \omega_{ab} = \left(\nabla_{[a}\xi_{b]}\right)^{\top}, \tag{25.7}$$

where the symbol $(\cdots)^{\top}$ indicates a projection in the $(e_{(1)}, e_{(2)})$ plane. The operators σ_{ab} and ω_{ab} are responsible for transforming vectors in the $(e_{(1)}, e_{(2)})$ plane. Consider now a set of neighboring null geodesics generated by ξ^a. Let ζ^a be a connecting vector, i. e., it is transverse to ξ^a and is Lie dragged along ξ^a: $\mathcal{L}_\xi\zeta^a = [\xi, \zeta]^a = \xi^b\nabla_b\zeta^a - \zeta^b\nabla_b\xi^a = 0$. We get the effect of the expansion, shear, and twist as operators in the $(e_{(1)}, e_{(2)})$ plane leading to the evolution of ζ^a in time

$$\begin{aligned}\dot{\zeta}_b^{\top} &:= \left(\xi^a\nabla_a\zeta_b\right)^{\top} = \zeta^a\left(\nabla_a\xi_b\right)^{\top} \\ &= \left(\sigma_{ab} + \omega_{ab} + \tfrac{1}{2}\Theta_{(\xi)}q_{ab}\right)\zeta^a.\end{aligned} \tag{25.8}$$

The effect of expansion, shear, and twist on ζ^a is illustrated in Fig. 25.1. A particularly important result is the Raychaudhuri equation which gives the time derivative of the expansion for affinely parameterized null geodesics

$$\frac{\mathrm{d}\Theta_{(\xi)}}{\mathrm{d}\lambda} = -\tfrac{1}{2}\Theta^2_{(\xi)} - \sigma_{ab}\sigma^{ab} + \omega_{ab}\omega^{ab} - T_{ab}\xi^a\xi^b. \tag{25.9}$$

This is a particular component of the Einstein field equations and a derivation and applications can be found in, e.g., [25.1, 3]. We shall have occasion to use it at various points during the course of this chapter.

Returning to the vector fields ℓ^a and n^a defined in (25.4), we see that they are manifestly orthogonal to the constant (v, r) spheres. Thus their expansions must involve area elements on these spheres, and it is in fact easy to calculate their expansions

$$\Theta_{(\ell)}(r) = \frac{r - 2M}{r^2}, \quad \Theta_{(n)}(r) = -\frac{2}{r}. \tag{25.10}$$

For spheres outside the black hole region, i. e., spheres with $r > 2M$, we see that $\Theta_{(\ell)} > 0$ and $\Theta_{(n)} < 0$. This is how a round sphere in flat space behaves. However, for spheres in the black hole region, we get both expansions to be negative. Such spheres are known as trapped surfaces and play a fundamental role in black hole theory and, in particular, in the singularity theorems. The

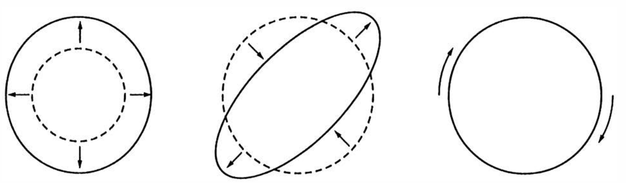


Fig. 25.1 From *left to right*, an illustration of expansion, shear, and twist, respectively, in the $(e_{(1)}, e_{(2)})$ plane transverse to ξ^a, as illustrated by the effect on circles in this plane. The expansion $\Theta_{(\xi)}$ is an isotropic expansion, the shear is an expansion, and contraction in orthogonal directions with the area being preserved, and the twist is a rotation

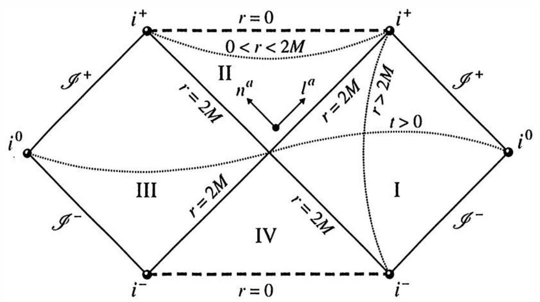


Fig. 25.2 Penrose–Carter conformal diagram for the extended Schwarzschild spacetime. See the text for details

spheres on the $r = 2M$ hyper-surface have $\Theta_{(\ell)} = 0$, $\Theta_{(n)} < 0$ and are called *marginally* trapped surfaces. Thus, we see that the $r = 2M$ hyper surface separates the region where the spherically symmetric trapped surfaces live and are a signature of a black hole spacetime. It is worth noting that the presence of trapped surfaces is not necessarily a signature of strong field gravity. The Riemann tensor (and thus the tidal force) is proportional to M/r^3. Thus at $r = 2M$, it is $\propto 1/M^2$, and large black holes have a correspondingly weaker curvature at their Schwarzschild radius. In this sense, trapped surfaces are a nonperturbative phenomenon in general relativity. An observer falling into a sufficiently large black will not notice anything out of the ordinary.

The role of the $r = 2M$ hyper-surface as the boundary of the region containing trapped surfaces (known as the trapped region) is typically not emphasized in standard textbooks on the subject. What is emphasized is instead the fact that starting from a point with $r < 2M$, there exist no timelike or null curves which can cross the $r = 2M$ hyper-surface. This is easiest to visualize in a Penrose–Carter conformal diagram shown in Fig. 25.2. This is a convenient way of visualizing the r–t part of the Schwarzschild metric. Just as we had extended the Schwarzschild metric in (25.1) across $r = 2M$ by using ingoing null coordinates, we can extend the metric of (25.3) further by going to double null coordinates (u, v) where

$$du = dt - \left(1 - \frac{2M}{r}\right) dr\,,$$
$$dv = dt + \left(1 - \frac{2M}{r}\right) dr\,. \tag{25.11}$$

Figure 25.2 is then obtained by performing a further rescaling of coordinates and a conformal transformation which brings infinity to a finite distance. Details can be found in [25.3, 6] or in other standard textbooks on the subject. The original Schwarzschild metric is region I in this diagram, while (25.3) corresponds to I and II. Regions III and IV are mirror images of I and II, respectively. In this figure, null curves are straight lines at 45°, i^0 is spatial infinity, i^+ is future timelike infinity and i^- is past timelike infinity. Future directed null curves in this figure end up either at the future singularity at $r = 0$ (marked with a dashed line) or at future null infinity labeled as $\mathcal{I}^+$ ($\mathcal{I}^-$ is past null infinity). It is also worth pointing out that if region I is defined to be the *outside world* so that ℓ^a is the outward pointing null normal (this is merely a matter of convention), then $\Theta_{(\ell)} < 0$ in regions II *and* III. However, region III has $\Theta_{(n)} > 0$ so that only region II has both expansions negative. Similarly, only region IV has both expansions positive.

It is then clear that no point in region II can be connected with region I (or III) by a causal curve and thus justifies the term *black hole* for region II. The boundary of this region occurs at $r = 2M$ and is called an event horizon. Note that the boundary of region IV also occurs at $r = 2M$ but region IV is a time-reversed version of II; it has the property that no future directed causal curve can stay within it. Thus, region IV is called a white hole. In the rest of this chapter, we shall only consider the portion of the $r = 2M$ hyper-surface which bounds the black hole, i. e., region II. In a physical gravitational collapse situation, the white hole (and region III) does not actually exist and is covered up by the matter fields which constitute the star; this will soon be seen explicitly when we study the Vaidya metric.

We thus have two different routes for describing a black hole: trapped surfaces versus event horizons (though one might suspect a priori that the two might be intimately related). Trapped surfaces seem to be more local than event horizons. To know that a particular null geodesic will not leave a particular region of spacetime, one might need to know the properties of the spacetime far away from the starting point of the geodesic. On the other hand, for trapped surfaces, in the Schwarzschild case we have just needed the computations of the expansions in (25.10) at a fixed value of (v, r). However, one should not forget that the computations of the expansions are not at just a single spacetime point, but are instead to be performed at all points over a sphere. This issue is irrelevant for spherically symmetric trapped surfaces in spherically symmetric spacetimes, but it is an important point that one cannot identify a trapped surface by examining only a part of it. For this reason, the trapped surface condition is said to be *quasi-local*.

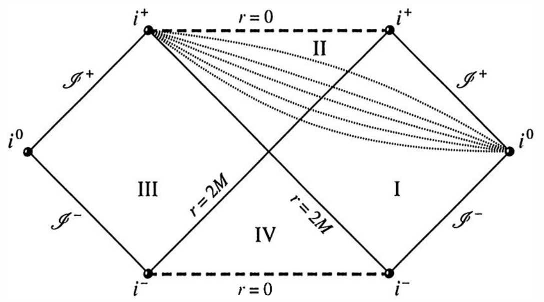


Fig. 25.3 A nonspherically symmetric spatial hyper-surface in the Schwarzschild spacetime depicted as a set of curves

In any case, for Schwarzschild, the two descriptions of the black hole region agree: the $r=2M$ hyper-surface is both the event horizon and also the boundary of the region of spacetime which contains trapped surfaces.

Visualizing nonspherically symmetric trapped surfaces is harder, even in a spacetime as simple as Schwarzschild. As in many numerical investigations in general relativity, let us try to locate such surfaces on three-dimensional spatial hyper-surfaces. The equation $\Theta_{(\ell)}=0$ turns into a minimization problem, and to a second-order elliptic equation in three-dimensional space (we shall have more to say on this matter later). The spatial hyper-surfaces depicted in Fig. 25.2 were all spherically symmetric, and thus a single curve in the Penrose diagram suffices for them. However, if we wish to depict nonspherically symmetric hyper-surfaces, we will need a collection of such curves, say one for each value of (θ,ϕ). An example is shown in Fig. 25.3. In this example, the spatial hyper-surface starts from i^+ on region III (but this detail is not important for our purposes and we could also have started from i^0 on the left edge of region III). What is important is that the spatial hyper-surfaces, or alternatively all the curves shown in Fig. 25.3, intersect the event horizon. As we shall see later, due to the somewhat nonintuitive properties of such null surfaces, the intersection of such a spatial hyper-surface with the $r=2M$ hyper-surface is still a marginally trapped surface though now a nonsymmetric one. More specifically, it turns out that ℓ_a is covariantly constant on the $r=2M$ surface so that $\nabla_a\ell_b$, projected onto the $r=2M$ surface, vanishes identically (see the discussion around (25.56) and in Sect. 25.4.4). This means that all closed cross-sections of the $r=2M$ surface are marginally trapped.

There would of course generally be nonspherically symmetric trapped surfaces on these spatial hyper-surfaces lying inside the marginally trapped one. Each spherically symmetric trapped and marginally trapped surface can be found by such a procedure. This construction clearly shows that there are many more nonsymmetric trapped surfaces than symmetric ones; each spherically symmetric hyper-surface can be deformed in an infinite number of ways and still contain trapped and marginally trapped surfaces.

We finally note that it is possible to come up with examples where part of the spatial hyper-surface extends arbitrarily close to the future singularity, but part of it is still outside the black hole region so that its intersection with the event horizon is not a complete sphere. There would then exist no marginally trapped surfaces on such a spatial hyper-surface [25.7].

25.2.2 The Vaidya Spacetime

As we have just seen, for a Schwarzschild black hole, all the natural definitions of the surface of a black hole agree. Thus, the $r=2M$ surface is both the boundary of the trapped region and also the event horizon. This has led to a widespread belief that the notion of a black hole and its surface is unambiguous. Matters are however not so simple in dynamical situations. Let us now look at what is perhaps the simplest example of a dynamical black hole, namely the spherically symmetric Vaidya spacetime [25.8].

The Vaidya solution is obtained by starting with the Schwarzschild metric in ingoing Eddington–Finkelstein coordinates of (25.3), and replacing the constant M by a nondecreasing function $M(v)$

$$\mathrm{d}s^2=-\left(1-\frac{2M(v)}{r}\right)\mathrm{d}v^2+2\mathrm{d}v\mathrm{d}r+r^2\mathrm{d}\Omega^2\,. \tag{25.12}$$

The stress energy tensor for this metric is

$$T_{ab}=\frac{\dot{M}(v)}{4\pi r^2}\nabla_a v\nabla_b v\,,\quad \text{where}\quad \dot{M}(v):=\frac{\mathrm{d}M(v)}{\mathrm{d}v}\,. \tag{25.13}$$

This represents the collapse of null dust to form a black hole and if $\dot{M}(v)\geq 0$, then T_{ab} satisfies the dominant energy condition. If we choose a mass function which is nonzero only for $v>v_0$ and constant after $v=v_1$, then we will have (portions of) Minkowski and Schwarzschild spacetimes for $v<v_0$ and $v>v_1$, respectively. The Penrose–Carter diagram for this spacetime

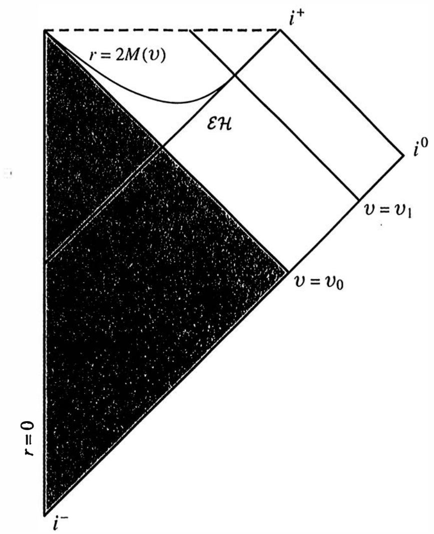


Fig. 25.4 Penrose–Carter conformal diagram for the Vaidya spacetime. The region shaded in *brown* is flat and the region in *light brown* is isomorphic to a portion of Schwarzschild. The event horizon is labeled $\mathcal{EH}$ and is seen to be distinct from the $r = 2M(v)$ surface. The two agree only in the final Schwarzschild portion

is shown in Fig. 25.4. A suitable set of null normals orthogonal to the constant (r, v) spheres are

$$\ell = \frac{\partial}{\partial v} + \frac{1}{2}\left(1 - \frac{2M(v)}{r}\right)\frac{\partial}{\partial r}\,, \qquad n = -\frac{\partial}{\partial r}\,, \tag{25.14}$$

and their expansions are, respectively, found to be

$$\Theta_{(\ell)}(v, r) = \frac{r - 2M(v)}{r^2}\,, \qquad \Theta_{(n)}(v, r) = -\frac{2}{r}\,. \tag{25.15}$$

Thus, in this case, there are no spherically symmetric trapped surfaces outside the $r = 2M(v)$ surfaces, and as in Schwarzschild, the spheres with $r = 2M(v)$ and fixed v are marginally trapped surfaces.

The event horizon is also not difficult to locate. Consider the outgoing null geodesics generated by the vector field ℓ above. Some of these geodesics will reach infinity, while others will terminate at the singularity. The event horizon is the boundary between the two cases. If we assume that the mass function reaches a finite final steady-state value M_∞, then the final black hole is a portion of Schwarzschild, and thus the condition $r = 2M_\infty$ defines the final state of the event horizon. Thus, we want to find the outgoing null geodesic generated by ℓ^a defined in (25.14) for which $r \to 2M_\infty$ when $v \to \infty$. Subject to this final state boundary condition, we need to solve

$$\frac{dr}{dv} = \frac{1}{2}\left(1 - \frac{2M(v)}{r}\right). \tag{25.16}$$

This is fairly easy to solve numerically for a generic mass function.

Let us now note a few properties of this event horizon. A typical case is shown in Fig. 25.4. First note that the event horizon extends to the flat region. A mortal observer in the flat region who has no way of knowing that the gravitational collapse will occur at some time in the future, might actually be living near an event horizon. The existence of the event horizon has really no consequences for any physical experiment or observations that the observer can conduct locally, and contrary to popular belief, the observer can cross the event horizon without feeling anything out of the ordinary. Furthermore, even an observer in the intermediate region $v_0 < v < v_1$, who can witness gravitational collapse occurring cannot know the true location of the event horizon. To illustrate this, consider Fig. 25.5. Here the mass function is nonzero for $v < v_0$ as before, however, there are two phases. The mass function first reaches a constant value at v_1 but restarts again at a later time v_2. The observer with $v < v_2$ cannot know the value of M_∞ and can thus never know the true location of the event horizon.

Examples of Nonsymmetric Trapped Surfaces. In contrast to the event horizon, the $r = 2M(v)$ surface seems to have the right properties. It bounds the region which has spherically symmetric trapped surfaces, it does not extend into the flat region, it grows only when $\dot{M} > 0$, it can be located quasi-locally, i. e., by checking the conditions for a trapped surface on a sphere, and it does not care about what happens to $M(v)$ at late times. However, nonspherically symmetric trapped surfaces are not so well behaved. While there are no marginally trapped surfaces which lie completely within the flat region, we shall see that portions of them can extend into the flat region.

One can try to find marginally trapped surfaces numerically. The standard procedure is to start with a particular spatial hyper-surface Σ, and to find a surface S in Σ for which $\Theta_{(\ell)} = 0$. Let h_{ab} be the Riemannian metric on Σ induced by g_{ab}. If the unit normal

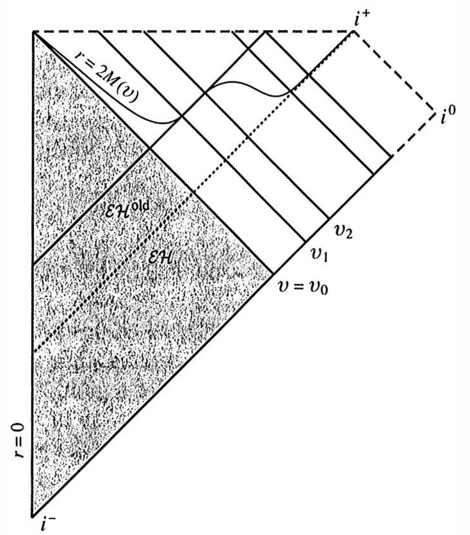


Fig. 25.5 Another Penrose–Carter conformal diagram for the Vaidya spacetime. The region shaded in *brown* is again the flat region. The mass function here has two phases where it is increasing. An observer with $v < v_2$ could in principle locate the *solid line* marked $\mathcal{EH}^{\text{old}}$ but will not know that the mass function will increase again and that the true event horizon is in fact given by the surface (indicated by a *dotted line* in this figure) marked $\mathcal{EH}$

on Σ is $\widehat{t}^a$, the unit spacelike normal to S on Σ is $\widehat{r}^a$ with $\ell^a = (\widehat{t}^a + \widehat{r}^a)/\sqrt{2}$ (see Fig. 25.6). With this choice, noting that the metric on S is $q_{ab} = h_{ab} - \widehat{r}_a\widehat{r}_b = g_{ab} + \widehat{t}_a\widehat{t}_b - \widehat{r}_a\widehat{r}_b$, the condition $\Theta_{(\ell)} = 0$ can be written as

$$\sqrt{2}\Theta_{(\ell)} = \sqrt{2}q^{ab}\nabla_a\ell_b = D_a\widehat{r}^a + K_{ab}\widehat{r}^a\widehat{r}^b - K = 0\,. \tag{25.17}$$

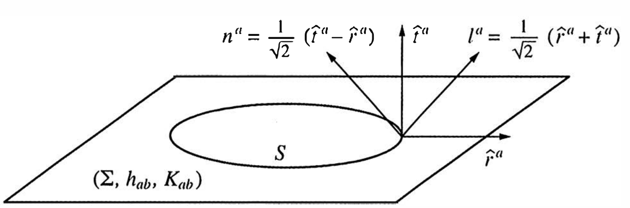


Fig. 25.6 A closed marginally outer trapped surface on a spatial Cauchy surface Σ with intrinsic metric h_{ab} and extrinsic curvature K_{ab}. The unit timelike normal to Σ is $\widehat{t}^a$ and the outward unit spatial normal to S is $\widehat{r}^a$. A particular choice of the out- and in-going null normals are $\frac{1}{\sqrt{2}}(\widehat{t}^a \pm \widehat{r}^a)$

Here $K_{ab} = h_a^c h_b^d \nabla_c \widehat{t}_d$ is the extrinsic curvature, K is its trace, and D is the derivative operator on Σ (we shall explain these concepts in more detail later in Sect. 25.3.2). Taking coordinates (r, θ, ϕ) on S and assuming that the surface is given by the equation $r = f(\theta, \phi)$, the above equation becomes a nonlinear second-order partial differential equation for f which can be solved numerically. Typical methods assume that the surface is *star-shaped*, i. e., every ray from the origin $r = 0$ intersects the surface exactly once; for a more complete description of this and other methods, we refer to [25.9].

We then choose a particular mass function and a Σ defined through a particular nonaxisymmetric time coordinate and attempt to locate surfaces with $\Theta_{(\ell)} = 0$. Let us review an illustrative result from a study reported in [25.10] (see also [25.11] for another such study). The particular mass function chosen corresponds to a short pulse of radiation

$$M(v) = \begin{cases} 0 & \text{for} \quad v \le 0\,, \\ \dfrac{M_0 v^2}{v^2 + W^2} & \text{for} \quad v > 0\,. \end{cases} \tag{25.18}$$

The parameter M_0 is the final mass and W determines the time-scale of the radiation pulse. We choose $M_0 = 1$ and $W = 0.1$. The nonspherically symmetric time coordinate is taken to be

$$\bar{t} = v - r(1 + \alpha \cos\theta)\,. \tag{25.19}$$

The constant α determines the degree of asymmetry, and we choose $\alpha = 10/11$. As in Fig. 25.3, spatial hyper-surfaces of constant $\bar{t}$ correspond to different sets of curves (one for each θ) in the Penrose–Carter diagram. Figure 25.7 shows the marginal surface found on the $\bar{t} = -0.3$ spatial hyper-surface. It shows the section in the x–z plane. The blue dotted line is the marginal surface, the green dashed line encloses the intersection of the flat region with the hyper-surface, and the *solid red curve* shows the intersection with the $r = 2M(v)$ surface (which is not a marginal surface). Thus, we see that the marginal surface extends in the flat region, though in this example it is planar with $\Theta_{(n)} = 0$ (so it is not, strictly speaking, a marginally trapped surface). The marginal surface is seen to be only partially inside the $r = 2M(v)$ surface.

The Trapped Region. Having looked at particular examples of trapped surfaces in a Vaidya spacetime, let us consider the trapped region, i. e., the portion of the

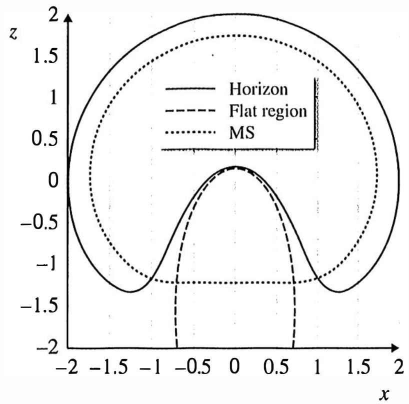


Fig. 25.7 A nonsymmetric closed surface in a particular Vaidya spacetime with $\Theta_{(\ell)} = 0$ and $\Theta_{(n)} \leq 0$. The solid red curve is the intersection of the spatial hyper-surface with the $r = 2M(v)$ surface, the *green dotted line* is the boundary of the flat region, and the *blue dashed line* is the marginal surface located on this spatial hyper-surface

manifold which contains trapped surfaces. The trapped region for the Vaidya spacetime can in fact be studied analytically. The starting point for this goes back to a conjecture by *Eardley* in 1998 [25.12]: *The outer boundary of the region containing outer trapped surfaces is the event horizon* (an outer trapped surface has $\Theta_{(\ell)} < 0$ and no restriction on $\Theta_{(n)}$). One the one hand, it is known that trapped surfaces cannot cross the event horizon. On the other hand, in dynamical situations like Vaidya, the event horizon is growing in area and its cross-sections are not marginally trapped. Thus, while the outer trapped surface might get arbitrarily close to the event horizon, the limiting process is not trivial. The Vaidya spacetime provides a relatively simple setting to study this phenomenon.

Recent works in this direction have been [25.10, 13–15]). For any point p in the flat portion of the black hole region of the Vaidya spacetime, *Ben-Dov* showed [25.13] that there exists an outer trapped surface S which contains p. This works even when p is arbitrarily close to the event horizon. In this case, most of the trapped surface actually lies inside the $r = 2M(v)$ surface in the far future where v is large. There is a narrow *tendril* which is almost null for a large portion, and yet stretches from the far future right down to the flat portion within the event horizon. This is precisely the kind of trapped surface whose existence was conjectured by *Eardley* in [25.12]. It is not clear that such highly nonsymmetric trapped surfaces would be present in typical spatial hyper-surfaces used in numerical relativity simulations, or in fact, whether the standard numerical methods currently employed would be able to locate such a surface even if it were present.

It is also possible to locate the boundary of closed future trapped surfaces (i. e., surfaces with $\Theta_{(\ell)} < 0$ and $\Theta_{(n)} < 0$) in Vaidya spacetimes. The first result was obtained by *Ben-Dov* [25.13], but the strongest results known at present are due to *Bengtsson* and *Senovilla* [25.14, 15]. Bengtsson and Senovilla have proved a number of results regarding the properties of trapped surfaces in spherically symmetric spacetimes, but here we shall only illustrate them by describing the past spacelike barrier for trapped surfaces in the Vaidya case. To this end, we need the following result [25.15, Theorem 4.1]: In a region $\mathcal{R}$ of a spacetime, let ξ^a be a future pointing hyper-surface orthogonal vector field so that $\xi_a = -F\nabla_a \bar{t}$ for some $F > 0$ and some $\bar{t}$ which increases to the future. If S is a future trapped surface which intersects $\mathcal{R}$ (but is not necessarily contained within $\mathcal{R}$), then S cannot contain a local minimum of $\bar{t}$ at points with $q^{ab}\mathcal{L}_\xi g_{ab} \geq 0$.

In a region $\mathcal{R}$ with a time coordinate such as $\bar{t}$, the significance of this results is that once a future trapped surface enters such a region with initially decreasing $\bar{t}$, (i. e., if the surface is initially *bending downward* in time), then $\bar{t}$ must continue decreasing. If $\mathcal{R}$ is bounded in the past by the event horizon, then clearly this result forces S to continue till it reaches the event horizon. Since S cannot cross the event horizon (or even touch it), it becomes clear that the region $\mathcal{R}$ cannot contain even portions of future trapped surfaces which are bending downward in time. In Vaidya, an appropriate ξ is the so-called Kodama vector

$$\xi^a = \left(\frac{\partial}{\partial v}\right)^a \Longrightarrow \xi_a = \nabla_a r - \left(1 - \frac{2M(v)}{r}\right)\nabla_a v\,. \tag{25.20}$$

Since $\xi_a\xi^a = -(1 - 2M(v)/r)$ it is clear that ξ^a is future directed to the past of the $r = 2M(v)$ surface. Furthermore, it is easy to check that

$$\mathcal{L}_\xi g_{ab} = 2\dot{M}(v)\ell_a\ell_b\,. \tag{25.21}$$

Thus, $q^{ab}\mathcal{L}_\xi g_{ab} \geq 0$ if $\dot{M} \geq 0$. The surfaces of constant $\bar{t}$, denoted by $\Sigma^{\bar{t}}$, are spherically symmetric and defined by

$$\frac{\mathrm{d}v}{\mathrm{d}r} = \left(1 - \frac{2M(v)}{r}\right)^{-1}\,. \tag{25.22}$$

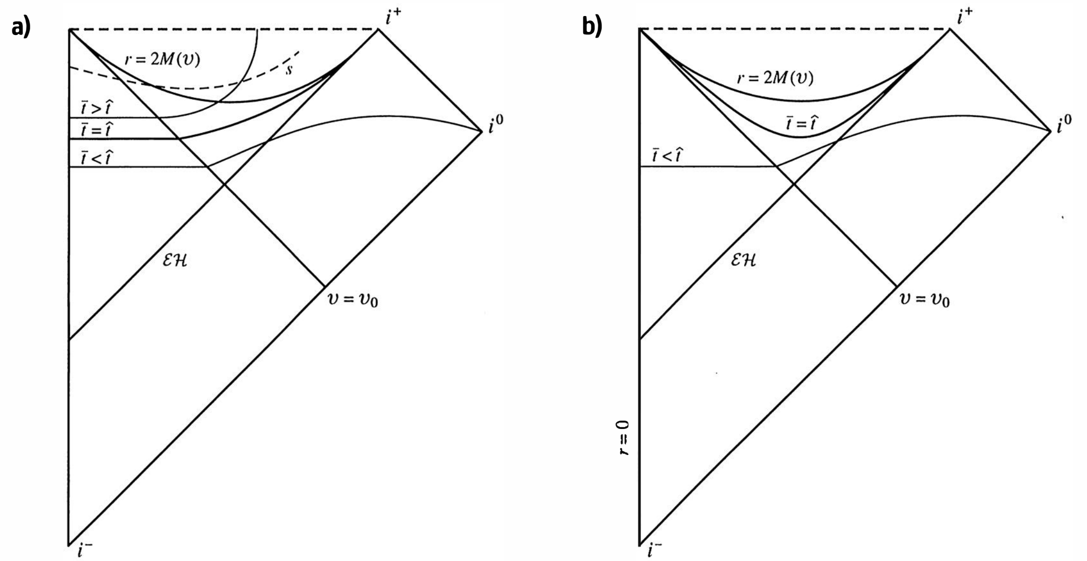


Fig. 25.8a,b Surfaces of constant Kodama time for a Vaidya spacetime with $\mu := \lim_{v \to 0} M(v)/v > 1/8$. In this case $\widehat{\Sigma}$ lies partially in the flat region. A future trapped surface which extends in the flat region is depicted as S. It required to enter the $r > 2M(v)$ region with increasing $\bar{t}$. Panel **(b)** is similar to panel **(a)**, but in this case $\widehat{\Sigma}$ does not extend into the flat region. This happens for $\mu \leq 8$. A similar picture also works when $M(v)$ asymptotes to a finite value for $v \to \infty$ (after [25.15, Fig. 15])

If $M(v)$ is constant for $v > v_1$, then there is a value of $\bar{t} = \hat{t}$ such that $\Sigma^{\bar{t}}$ coincides with the event horizon for $v > v_1$. Alternatively, if $M(v)$ asymptotes to a constant value, then there is a value of $\bar{t} = \hat{t}$ such that $\Sigma^{\bar{t}}$ also asymptotes to the event horizon. In both these cases, let us denote these $\Sigma^{\bar{t}}$ as $\widehat{\Sigma}$. In the flat region, $\Sigma^{\bar{t}}$ are horizontal lines in the Penrose diagrams. The behavior of $\widehat{\Sigma}$ depends sensitively on $M(v)$. In particular, it depends on the quantity $\mu := \lim_{v \to 0} M(v)/v$. When $\mu > 1/8$, it is shown in [25.15] that $\widehat{\Sigma}$ will enter the flat region, but in other cases it may not. These cases are depicted in Fig. 25.8 along with $\Sigma^{\bar{t}}$ for some other values of $\bar{t}$. Any trapped surface which extends outside the $r = 2M(v)$ surface must do so with increasing $\bar{t}$ and it must never *bend* downward in time. If it does not do so, then the above result shows that it must continue downward till it hits the event horizon where it must cease to be smooth if the condition $\Theta_{(\ell)} < 0$ is to be maintained (the event horizon is expanding and thus has positive expansion). It cannot terminate smoothly because this would imply the existence of a point on the trapped surface where $\bar{t}$ is a local minimum.

A little thought then shows that $\widehat{\Sigma}$ must then be a past barrier which future trapped surfaces cannot cross. Furthermore, it is also clear that there cannot be any compact future trapped surface contained entirely in the region to the past of $\widehat{\Sigma}$, and bounded between $\widehat{\Sigma}$ and the event horizon; if there were, again there would have to be a point on the trapped surface where $\bar{t}$ is a local minimum. The boundary of the region containing trapped surfaces has thus been located. This boundary $\widehat{\Sigma}$ is of course spherically symmetric (as one can prove on general grounds). However, it is not foliated by marginally trapped surfaces. Thus, as with the limit of outer trapped surfaces to the event horizon, it is clear that the limit of marginally trapped surfaces to this boundary cannot be smooth. $\widehat{\Sigma}$ is not a quasi-local object; it is as nonlocal as the horizon. Moreover, as far as we know, it does not have any features which might distinguish it as a black hole horizon.

Lessons from Spherical Symmetry. After this extensive discussion of the Schwarzschild and the Vaidya examples, let us summarize the situation in spherical symmetry. We have seen the complications that can arise from having to consider nonspherically symmetric trapped surfaces in dynamical situations. In

Schwarzschild there are no major surprises and any trapped surface can extend right up to the event horizon. This is not so in Vaidya; the obvious generalization of the $r = 2M$ surface from Schwarzschild does not coincide with the event horizon. We need nonspherically symmetric trapped surfaces to *fill the gap* between the $r = 2M(v)$ surface and the event horizon. It however turns out to be possible to study the trapped region in detail and to obtain a fairly complete understanding of where trapped surfaces can (and cannot) occur. There turns out to be a difference between future- and outer-trapped surfaces (i. e., whether or not we consider the $\Theta_{(n)} < 0$ condition). Outer trapped surfaces can extend all the way to the event horizon, but understanding how such surfaces limit to the event horizon is subtle; there is a separate past barrier for future trapped surfaces which is distinct from the event horizon.

What does this study tell us about nonspherically symmetric dynamical spacetimes? We note that the essential complication here, as noted by *Eardley* [25.12], is not that the black hole is nonspinning etc. Rather, the problem is to consider trapped surfaces which do not share the symmetry of the spacetime and to understand what happens to them as they are deformed toward the event horizon which will generally have positive expansion in dynamical situations. We might still expect outer trapped surfaces to extend to the event horizon in a similar fashion, but there is, in general, probably no separate barrier for future trapped surfaces.

The next step in this chapter will be to study how marginally trapped surfaces evolve in time and under general deformations, and this leads to the subject of quasi-local horizons. However, before doing so, we shall first formalize many of the ideas introduced in this section with some general definitions and results.

25.3 General Definitions and Results: Trapped Surfaces, Stability and Quasi-local Horizons

25.3.1 Event Horizons

The surface of a black hole is traditionally defined in terms of an event horizon which is the boundary of the region from where massive or mass-less particles can reach the outside world. The formal definition is however more involved, with the main difficulty being in how the *outside world* is to be defined. It is worthwhile to briefly sketch the various technical ingredients that go into the precise formal definition, if only to highlight once again the truly global nature of event horizons; see, e.g., [25.1, 3] for details.

This requires one to attach future and past null infinity $\mathcal{I}^{\pm}$ as boundaries to the physical spacetime and to consider the causal past of $\mathcal{I}^{+}$. This causal past is the region of spacetime $\mathcal{R}$ from which causal signals can escape to infinity and represents the *outside world*. The future boundary of $\mathcal{R}$ is the event horizon. In Fig. 25.2, the region $\mathcal{R}$ is the union of regions I, II, and III, the black hole is of course region II and the portion of the $r = 2M$ surface which divides region II from I and III is the event horizon. A little thought shows that in order for this notion to capture the physical idea we have in mind, it is necessary to ensure that $\mathcal{I}^{+}$ is complete in an appropriate sense. For example, if we were to look at the causal past of just a portion of $\mathcal{I}^{+}$ even for Minkowski space, we would erroneously conclude the presence of a black hole in Minkowski space [25.16]. Similarly for Schwarzschild, we could end up with the wrong location of the event horizon if we looked at only a portion of $\mathcal{I}^{+}$ (see again Fig. 25.2).

Since we need to construct a complete $\mathcal{I}^{+}$ and its global past in this definition, it is clear that the event horizon is a very global and teleological notion. There may in fact be an event horizon forming and growing in the room you are reading this chapter right now, because of possible events which might occur a billion years from now. An example is an observer in the flat region of the Vaidya spacetime in Fig. 25.4. This discussion also illustrates that formally, the notion of an event horizon cannot be used in a cosmological spacetime such as the one we inhabit, since it fails to be asymptotically flat.

In practice, the principle of calculating event horizons in numerical simulations is to start from an educated guess at late times, and to integrate a null geodesic or a null surface backward in time [25.9, 17]. The basis for these methods is the fact that when we integrate forward in time, a small initial error will cause the null geodesic or surface to diverge exponentially from the true solution and end up either in the singularity or at infinity. This implies that by integrating backward from

even a reasonable guess at late times one will converge to the true solution exponentially.

25.3.2 Trapped Surfaces

Let us begin with the standard definitions of the first and second fundamental form of a smooth nondegenerate submanifold $\widetilde{\mathcal{M}}$ embedded in a spacetime $(\mathcal{M}, g_{ab})$; our discussion mostly follows [25.4]. The first fundamental form is just the induced metric metric h_{ab} on $\widetilde{\mathcal{M}}$ so that for any vectors X^a and Y^b tangent to $\widetilde{\mathcal{M}}$

$$h_{ab}X^aY^b := g_{ab}X^aY^b\,. \tag{25.23}$$

If h_{ab} is nondegenerate, we can decompose the tangent space T_pM at any point $p \in \widetilde{\mathcal{M}}$ as

$$T_pM = T_p\widetilde{\mathcal{M}} \oplus T_p^{\perp}\widetilde{\mathcal{M}}\,, \quad T_p\widetilde{\mathcal{M}} \cap T_p^{\perp}\widetilde{\mathcal{M}} = \{0\}\,, \tag{25.24}$$

where $T_p\widetilde{\mathcal{M}}$ is the tangent space to $\widetilde{\mathcal{M}}$ and the subspace $T_p^{\perp}\widetilde{\mathcal{M}}$ is normal to it. Thus, an arbitrary nonvanishing vector field ξ defined at points of $\widetilde{\mathcal{M}}$ can be split uniquely into a tangential and normal part

$$\xi = \xi^{\top} + \xi^{\perp} \quad \text{where} \quad \xi^{\perp}\cdot X = 0\,, \tag{25.25}$$

for any vector X tangent to $\widetilde{\mathcal{M}}$. Then, for any X, Y tangent to $\widetilde{\mathcal{M}}$ we have

$$\nabla_X Y = (\nabla_X Y)^{\top} + (\nabla_X Y)^{\perp}\,. \tag{25.26}$$

The intrinsic covariant derivative on $\widetilde{\mathcal{M}}$ is then defined as $D_X Y := (\nabla_X Y)^{\top}$. The second fundamental tensor Π is an operator which takes two vectors tangent to $\widetilde{\mathcal{M}}$ and produces a vector in the normal-subspace: $T_p\widetilde{\mathcal{M}} \times T_p\widetilde{\mathcal{M}} \to T_p^{\perp}\widetilde{\mathcal{M}}$

$$\Pi(X, Y) := (\nabla_X Y)^{\perp}\,. \tag{25.27}$$

It is easy to show that D defined this way is a legitimate derivative operator, and that the second fundamental form is symmetric: $\Pi(X, Y) = \Pi(Y, X)$. The symmetry of Π is especially easy

$$\begin{aligned}\Pi(X, Y) - \Pi(Y, X) &= (\nabla_X Y - \nabla_Y X)^{\perp} = [X, Y]^{\perp}\\ &= 0\,.\end{aligned} \tag{25.28}$$

In the last step we have used the Frobenius theorem which says that for a smooth submanifold $\widetilde{\mathcal{M}}$, if X, Y are tangent to $\widetilde{\mathcal{M}}$, then so is their commutator.

When $\widetilde{\mathcal{M}}$ is a hyper-surface, i. e., when it has codimension 1, then N_p is one-dimensional and spanned by the unit-normal N^a. Thus we can define the *extrinsic curvature* K_{ab} via

$$\Pi(X, Y)^c := -(K_{ab}X^aY^b)N^c\,. \tag{25.29}$$

The symmetry of Π implies that $K_{ab} = K_{ba}$. The most important case for us is however when the submanifold S is two-dimensional and spacelike; the first fundamental form, denoted by q_{ab} here, is a Riemannian metric on S. We can again make the decomposition $T_p\mathcal{M} = T_pS \oplus T_p^{\perp}S$. The normal space $T_p^{\perp}S$ is a $1+1$-dimensional Minkowski space. It will be convenient to choose two null vectors ℓ and n to span $T_p^{\perp}S$. We are of course free to rescale ℓ and n independently by scalars, but we choose to use a normalization $\ell\cdot n = -1$ which cuts down the rescaling freedom to

$$\ell \to A\ell\,, \quad n \to A^{-1}n\,. \tag{25.30}$$

We will always choose (ℓ, n) to be future directed and ℓ and n as outward and inward pointing, respectively. The *mean curvature* vector K^a is the trace of the second fundamental form

$$K^c := q^{ab}\Pi_{ab}{}^c = -\Theta_{(n)}\ell^c - \Theta_{(\ell)}n^c\,. \tag{25.31}$$

The coefficients appearing here turn out to be precisely the expansions $\Theta_{(\ell,n)}$ discussed earlier. Under a rescaling of the kind in (25.30), K^a remains invariant.

The different kinds of trapped surfaces correspond to properties of K^a. Two useful definitions are:

- *Future trapped surface* ($\Theta_{(\ell)} < 0$, $\Theta_{(n)} < 0$): K^a is timelike and future-directed.
- *Marginally future trapped surface* ($\Theta_{(\ell)} = 0$, $\Theta_{(n)} < 0$): K^a is null and future-directed.

Furthermore, we shall consider only *closed* surfaces; this is an important condition as it, among other things, excludes trivial planar surfaces in flat space.

It is also useful to remark on the physical significance of the condition $\Theta_{(n)} < 0$. This condition holds for round spheres in flat space and also, as we have seen, for round spheres in the trapped region of Schwarzschild (see (25.10)). However, this may not be true for nonspherically symmetric trapped surfaces even in Schwarzschild. Furthermore, there are

explicit black hole solutions found by *Geroch* and *Hartle* [25.18] which represent black holes distorted by surrounding matter fields which do not satisfy $\Theta_{(n)} < 0$ at all points on the horizon (but its average value over the horizon is still negative). In fact, for a number of key results that we shall mention below, it is not necessary to impose $\Theta_{(n)} < 0$. Surfaces with the vanishing outward expansion $\Theta_{(\ell)} = 0$, with no restrictions on the sign of $\Theta_{(n)}$ will be called *marginally outer trapped surfaces*, usually abbreviated to MOTS. A surface with $\Theta_{(\ell)} < 0$ and no condition on $\Theta_{(n)}$ will be said to be outer trapped. We have seen in the Vaidya example that there are differences in the location of trapped and outer trapped surfaces.

The collection of closed future-trapped surfaces form the trapped region of the spacetime. More precisely, the trapped region $\mathcal{T}$ consists of spacetime points which lie on a closed future-trapped surface. The boundary $\mathcal{B}$ of $\mathcal{T}$ is the trapping boundary. It is also common to consider these concepts restricted to a spacelike surface Σ, typically a Cauchy surface in an initial value problem set-up. The trapped $\mathcal{T}^{\Sigma}$ region on Σ is the set of points on Σ which lie on a closed future-trapped surface contained entirely on Σ. The boundary of $\mathcal{T}^{\Sigma}$ is denoted by $\mathcal{B}^{\Sigma}$, and each connected component of $\mathcal{B}^{\Sigma}$ is called an *apparent horizon* if it is the outermost such boundary on Σ. Since $\mathcal{B}^{\Sigma}$ excludes trapped surfaces not contained in Σ, it is clear that $\mathcal{B}^{\Sigma} \subset \mathcal{B} \bigcap \Sigma$. It is important to not confuse the trapped region $\mathcal{T}$ or its boundary $\mathcal{B}$ with the black hole region $\mathcal{B}$ defined in Sect. 25.3.1 and the event horizon. It can be shown [25.19] that with some additional regularity conditions, each connected component of the apparent horizon $\mathcal{B}^{\Sigma}$ is actually a closed marginally trapped surface. The same result was proved in [25.20] with the regularity assumptions removed. In practice, this fact is what is used to locate the apparent horizon in numerical simulations.

25.3.3 The Stability of Marginally Trapped Surfaces, Trapping, and Dynamical Horizons

Let us now go beyond individual trapped/marginally trapped surfaces and look at their time evolutions. Consider a region of spacetime foliated by smooth spacelike surfaces Σ^t depending on a real time parameter t. Start with initial data (the first and second fundamental forms) at $t = 0$ and evolve it using the Einstein and matter field equations. This way, we obtain a solution to the field equations locally in time near Σ^0. The first question we wish to address is: If Σ^0 contains an MOTS S_0, does it persist under time evolution and does it evolve smoothly? If it does evolve smoothly, then the union of all the MOTS S^t will form a smooth 3-surface $\mathcal{H}$ which we shall call a *marginally trapped tube* (MTT). A related question is then: How does $\mathcal{H}$ depend on the foliation Σ^t? If we start with the same Σ^0 but choose the surfaces differently for $t > 0$ (still requiring Σ^t to form a smooth foliation), then will we still end up with a smooth MTT $\mathcal{H}'$? If it exists, is it different from $\mathcal{H}$? In numerical simulations, it is found that the apparent horizon can evolve discontinuously. Can this be understood analytically? It turns out that the answers to the above questions are intimately connected with the *stability* of S_0 with respect to variations on Σ^0. To this end, we need to define the notion of the *geometric variation* of a 2-surface S which is embedded in a spacetime $\mathcal{M}$ [25.21–24]. Such a variation is a very general concept; it includes evolving in time, and an evolution following Einstein equations is a particular case. A smooth variation of a submanifold S is defined as a one-parameter family of surfaces S_λ (where λ is a real parameter and takes values in some interval $(-\epsilon, \epsilon)$) such that: S_0 is identical to S, each S_λ is a smooth surface, and each point on S moves on a smooth curve as λ is varied. We can then define a vector field q^a as the tangent to these curves. This is depicted in Fig. 25.10. We could also perform the variation along null directions or spatial directions in a given spacetime and of course, the variations do not need to be uniform on S and different

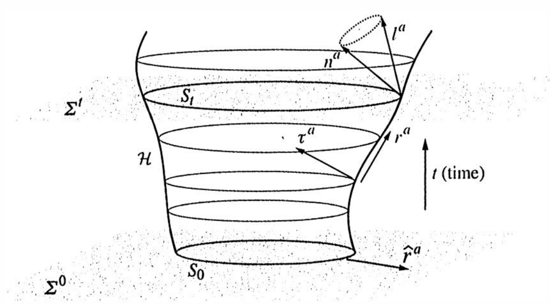


Fig. 25.9 The evolution of an MOTS in time. Start at $t = 0$ with an MOTS S_0 on a spatial hyper-surface Σ^0. Evolve the data on Σ^0 using the Einstein equations. If the evolution of the MOTS is smooth in time, then S will evolve to an MOTS S_t on Σ^t at time $t > 0$. The collection of all the S_t will then form a smooth 3-surface $\mathcal{H}$. We shall see that $\mathcal{H}$ will usually be spacelike so that the future null vectors ℓ^a and n^a orthogonal to S_t (and the future null cone) will point *inward* on $\mathcal{H}$. The vector $\hat{r}^a$ is the unit outward normal to S_t on Σ^t

points on S can move at very different speeds depending on q^a.

If we have a relevant geometric quantity on S, we can compute it on each S_ϵ and differentiate it with respect to λ, and the derivative is called the variation of that geometric quantity. If $\mathcal{O}_\lambda$ is such a geometric quantity (e.g., the expansion $\Theta_{(\ell)}$ on each S_λ), then we shall define $\delta_q\mathcal{O} := \partial_\lambda \mathcal{O}_\lambda|_{\lambda=0}$. It is also important to keep in mind that for a function f, in general the variation is not linear: $\delta_{fq}\mathcal{O} \neq f\delta_q\mathcal{O}$.

For an MOTS defined on a spatial hyper-surface Σ, the relevant variation is along $\widehat{r}^a$, the unit normal to S. The stability of S is meant to capture the idea that if S is deformed outward, it becomes untrapped, and an inward deformation leads to $\Theta_{(\ell)} < 0$. No condition on $\Theta_{(n)}$ is assumed. More precisely, the MOTS S is said to be *stably outermost* if there exists a function $f \geq 0$ on S such that $\delta_{f\widehat{r}}\Theta_{(\ell)} \geq 0$. S is *strictly stably outermost* if in addition $\delta_{f\widehat{r}}\Theta_{(\ell)} \neq 0$ somewhere on S.

With this background, we can state the following result [25.23, 24]: If S_0 is a smooth MOTS on Σ^0, and S_0 is strictly stably outermost on Σ^0, then S_0 evolves smoothly into smooth MOTSs S_t on Σ^t at time t at least for sufficiently small (but nonzero) t. Furthermore, the union of the S_t forms a smooth 3-surface which we shall call $\mathcal{H}$. This holds at least as long as the S_t continue to remain strictly stable outermost. In addition, if $G_{ab}\ell^a\ell^b > 0$ somewhere on S or if ℓ^a has nonvanishing shear somewhere on S, then $\mathcal{H}$ is spacelike. More generally, if the null energy condition holds then $\mathcal{H}$ is either spacelike or null.

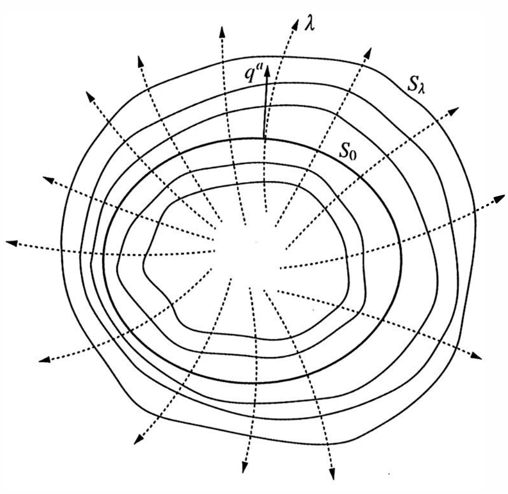


Fig. 25.10 The variation of a surface S viewed as a set of surfaces S_λ. The initial surface is S_0. Each point on S moves along a smooth curve and q^a is the tangent to these curves

These results answer several of the questions raised at the beginning of this subsection. If we were to start with the same Σ_0 but choose different ones at later times, the MTT would still exist at least for some small time interval since the stability condition still holds at $t = 0$. However, $\mathcal{H}'$ and $\mathcal{H}$ would not necessarily coincide. The jumps that are observed in numerical simulations are because of the outermost condition: while an individual MTT can continue to evolve smoothly, a new MTT can appear further outward. While not all MOTSs satisfy the stability condition, in all the cases that the author is aware of in numerical simulations, the MTTs continue to evolve smoothly. This suggests that the mathematical results could be strengthened.

Having seen that there is a physically interesting class of MOTSs which evolve smoothly in time, it is reasonable to impose additional conditions on an MTT to capture the fact that they are black hole horizons. The first is the notion of a *trapping horizon* defined by *Hayward* [25.25, 26] in 1994: A *future-outer-trapping horizon* (FOTH) is a smooth three-dimensional manifold $\mathcal{H}$ foliated by compact 2-manifolds S_t such that:

1. The surfaces S are marginally trapped surfaces ($\Theta_{(\ell)} = 0$, $\Theta_{(n)} < 0$),
2. The directional derivative of $\Theta_{(\ell)}$ along n^a vanishes: $\mathcal{L}_n\Theta_{(\ell)} < 0$.

Historically, this was the first systematic definition of a quasi-local horizon and it played a key role in spurring further developments. At the time, it was not known whether an MOTS would evolve smoothly in time and if trapped surfaces would limit smoothly to the boundary of the trapping region (as we have seen in the Vaidya example, they in fact do not and this possibility was recognized in [25.25]). Using this definition, Hayward showed that $\mathcal{H}$ would be generically spacelike and is null if and only if the matter flux and the shear of ℓ^a vanish identically (the result from [25.23, 24] quoted previously is stronger). Hayward also showed that it was possible to assign a surface gravity to the black hole, and to have versions of the laws of black hole mechanics applicable to $\mathcal{H}$. The second condition seems similar to the stability condition defined above, but in fact the two may not necessarily agree. A FOTH requires implicitly that ℓ^a and n^a have been extended smoothly in a neighborhood of S and does not refer to the variation of S.

A *dynamical horizon* [25.27, 28] is defined similarly, but it is designed explicitly for the case when the

horizon is spacelike: A smooth three-dimensional manifold $\mathcal{H}$ is said to be a *dynamical horizon* (DH) if it is foliated by compact two-dimensional manifolds S_t such that:

1. The surfaces S_t are marginally trapped ($\Theta_{(\ell)} = 0$, $\Theta_{(n)} < 0$).
2. $\mathcal{H}$ is spacelike.

The first condition is the same as for a FOTH, but the second condition specifies a priori the causal nature of $\mathcal{H}$ without any additional conditions on fields transverse to $\mathcal{H}$. Even though the definitions of a FOTH and a DH are similar, neither implies the other. We shall return to dynamical and trapping horizons later in Sect. 25.5, but before that, we shall discuss the equilibrium case when the MTT is null in some detail.

25.4 The Equilibrium Case: Isolated Horizons

Having discussed some general properties of trapped surfaces and their evolutions, we shall now consider a special case that is nevertheless of significant interest for various applications, namely when the black hole is in equilibrium with its surroundings and there is no matter or radiation falling into it. Since the work of the late 1990s and the subsequent years, our understanding of this special case has matured and we can now consider it be well understood. All the well-known globally stationary solutions in four-dimensions, i. e., the Kerr–Newman black holes, are included in this analysis. Also included are the dynamical cases when the black hole is itself in equilibrium but when the time dependent fields are relatively far away from the black hole. An example is a system of binary black holes when the separation of the black holes is much larger than either of the masses; to a very good approximation each black hole can be treated as being in equilibrium with its surroundings locally. In the same system, once the two black holes have coalesced, the final black hole will soon reach equilibrium after it's ringdown phase. The assumption of global stationarity obviously does not hold in these cases. Each of these cases is well modeled by the framework of isolated horizons that we shall now describe. We shall start with a prototypical example namely, the structure of the horizon of a Kerr black hole. This shall be followed by a general definitions and a summary of some key results. We shall conclude with two applications: the near horizon geometry and the mechanics of isolated horizons. It will be convenient to use the Newman–Penrose formalism in this discussion, and therefore we start with a short digression to discuss this formalism.

We note that strictly speaking, it is not essential to use this formalism for many of the results that we will discuss later. However, it does prove to be very useful in a number of cases. For our purposes, we discuss it here because it makes calculations very explicit and is thus useful for pedagogically; there are no tensor indices and all geometric quantities are expressed in terms of scalar functions which have clear geometric and physical interpretations. Since we wish to start with the Kerr black hole as an example, this formalism is particularly useful for studying its properties explicitly.

25.4.1 The Newman–Penrose Formalism

The Newman–Penrose formalism [25.29, 30] is a tetrad formalism where the tetrad elements are null vectors, which makes it especially well suited for studying null surfaces. See [25.2, 5, 6] for pedagogical treatments (note that these references take the spacetime metric to have a signature of $(+---)$ which is different from ours). Start with a null tetrad $(\ell, n, m, \bar{m})$ where ℓ and n are real null vectors, m is a complex null vector and $\bar{m}$ its complex conjugate. The tetrad is such that $\ell \cdot n = -1$, $m \cdot \bar{m} = 1$, with all other inner products vanishing. The spacetime metric is thus given by

$$g_{ab} = -\ell_a n_b - n_a \ell_b + m_a \bar{m}_b + \bar{m}_a m_b \,. \qquad (25.32)$$

Directional derivatives along the basis vectors are denoted as

$$D := \ell^a \nabla_a \,, \quad \Delta := n^a \nabla_a \,, \\ \delta := m^a \nabla_a \,, \quad \bar{\delta} := \bar{m}^a \nabla_a \,. \qquad (25.33)$$

We shall see that the Newman–Penrose formalism employs almost all the Greek symbols, and therefore often leads to conflicts in notation. An example is the symbol Δ which is used for both the directional derivative along n^a and the isolated horizon itself; we will soon have other such examples. Hopefully the notation should be clear from the context.

The components of the connection are encoded in 12 complex scalars, the spin coefficients, defined via

the directional derivatives of the tetrad vectors:

$$D\ell = (\epsilon+\bar\epsilon)\ell - \bar\kappa m - \kappa\bar m\,, \tag{25.34a}$$
$$Dn = -(\epsilon+\bar\epsilon)n + \pi m + \bar\pi m\,, \tag{25.34b}$$
$$Dm = \bar\pi\ell - \kappa n + (\epsilon-\bar\epsilon)m\,, \tag{25.34c}$$
$$\Delta\ell = (\gamma+\bar\gamma)\ell - \bar\tau m - \tau\bar m\,, \tag{25.34d}$$
$$\Delta n = -(\gamma+\bar\gamma)n + \nu m + \bar\nu\bar m\,, \tag{25.34e}$$
$$\Delta m = \bar\nu\ell - \tau n + (\gamma-\bar\gamma)m\,, \tag{25.34f}$$
$$\delta\ell = (\bar\alpha+\beta)\ell - \bar\rho m - \sigma\bar m\,, \tag{25.34g}$$
$$\delta n = -(\bar\alpha+\beta)n + \mu m + \bar\lambda\bar m\,, \tag{25.34h}$$
$$\delta m = \bar\lambda\ell - \sigma n + (\beta-\bar\alpha)m\,, \tag{25.34i}$$
$$\bar\delta m = \bar\mu\ell - \rho n + (\alpha-\bar\beta)m\,. \tag{25.34j}$$

While this kind of expansion may seem to some like a backward step to the days before efficient tensor notation was developed, it is actually very convenient in cases where null vectors and null surfaces are involved. The use of complex functions is also efficient because it cuts down the number of quantities by half.

Many of the spin coefficients have a clear geometric meaning. First note that (25.34a) implies that ℓ^a is geodesic if and only if $\kappa = 0$. Furthermore, it will be affinely parameterized if $\epsilon+\bar\epsilon = 0$. Similarly, from (25.34e), we see that n^a is geodesic if and only if $\nu = 0$ and it is affinely parameterized if $\gamma+\bar\gamma = 0$. Important for us in particular are the coefficients ρ and σ which are related to the expansion, shear and twist defined earlier in Sect. 25.2.1. Let ℓ^a be tangent to a null geodesic congruence, let it be affinely parameterized, and let ζ^a be a connecting vector for the congruence (it is transverse to ℓ^a and satisfies $[\ell,\zeta] = 0$). As in (25.8), we need to look at $\nabla_a\ell_b$ projected in the $(m,\bar m)$ plane. Note that in this plane, the metric is $q^{ab} = 2m^{(a}\bar m^{b)}$, the antisymmetric area form is ${}^2\epsilon = 2im^{[a}m^{b]}$, and the two-dimensional space of symmetric trace-free second-rank tensors is spanned by m^am^b and $\bar m^a\bar m^b$. From the above definitions, it is easy to show that

$$\Theta_{(\ell)} = q^{ab}\nabla_a\ell_b = m^a\bar\delta\ell_a + \bar m^a\delta\ell_a = -2\,\mathrm{Re}\,\rho\,, \tag{25.35}$$

$$m^{[a}\bar m^{b]}\nabla_a\ell_b = \mathrm{Im}\,\rho\,, \quad m^am^b\nabla_a\ell_b = -\sigma\,. \tag{25.36}$$

Thus, we see that $-2\,\mathrm{Re}\,\rho$ is the expansion, $\mathrm{Im}\,\rho$ is related to the twist, and σ is the shear of ℓ. Similarly, the real and imaginary parts of μ gives the expansion and twist of n^a, while λ yields its shear. It is also easy to verify that

$$[m,\bar m]^a = (\bar\mu-\mu)\ell^a + (\bar\rho-\rho)n^a + (\alpha-\bar\beta)m^a + (\beta-\bar\alpha)\bar m^a\,. \tag{25.37}$$

Thus, using the Frobenius theorem, we see that m and $\bar m$ can be integrated to yield a smooth surface if ℓ^a and n^a are twist free. Furthermore, since the projection of δm is determined by $\beta-\bar\alpha$, it is clear that this determines the connection, and thus the curvature of this surface.

Since the null tetrad is typically not a coordinate basis, the above definitions of the spin coefficients lead to nontrivial commutation relations

$$(\Delta D - D\Delta)f = (\epsilon+\bar\epsilon)\Delta f + (\gamma+\bar\gamma)Df - (\bar\tau+\pi)\delta f - (\tau+\bar\pi)\bar\delta f\,, \tag{25.38a}$$

$$(\delta D - D\delta)f = (\bar\alpha+\beta-\bar\pi)Df + \kappa\Delta f - (\bar\rho+\epsilon-\bar\epsilon)\delta f - \sigma\bar\delta f\,, \tag{25.38b}$$

$$(\delta\Delta - \Delta\delta)f = -\bar\nu Df + (\tau-\bar\alpha-\beta)\Delta f + (\mu-\gamma+\bar\gamma)\delta f + \bar\lambda\bar\delta f\,, \tag{25.38c}$$

$$(\bar\delta\delta - \delta\bar\delta)f = (\bar\mu-\mu)Df + (\bar\rho-\rho)\Delta f + (\alpha-\bar\beta)\delta f - (\bar\alpha-\beta)\bar\delta f\,. \tag{25.38d}$$

The Weyl tensor C_{abcd} breaks down into five complex scalars

$$\Psi_0 = C_{abcd}\ell^am^b\ell^cm^d\,, \tag{25.39a}$$
$$\Psi_1 = C_{abcd}\ell^am^b\ell^cn^d\,, \tag{25.39b}$$
$$\Psi_2 = C_{abcd}\ell^am^b\bar m^cn^d\,, \tag{25.39c}$$
$$\Psi_3 = C_{abcd}\ell^an^b\bar m^cn^d\,, \tag{25.39d}$$
$$\Psi_4 = C_{abcd}\bar m^an^b\bar m^cn^d\,. \tag{25.39e}$$

Similarly, the Ricci tensor is decomposed into four real and three complex scalars Φ_{ij}:

$$\begin{aligned}\Phi_{00} &= \tfrac{1}{2}R_{ab}\ell^a\ell^b\,,\\ \Phi_{11} &= \tfrac{1}{4}R_{ab}(\ell^an^b + m^a\bar m^b)\,,\\ \Phi_{22} &= \tfrac{1}{2}R_{ab}n^an^b\,,\\ \Lambda &= \frac{R}{24}\,,\end{aligned} \tag{25.40a}$$

$$\begin{aligned}\Phi_{01} &= \tfrac{1}{2}R_{ab}\ell^am^b\,,\\ \Phi_{02} &= \tfrac{1}{2}R_{ab}m^am^b\,,\\ \Phi_{12} &= \tfrac{1}{2}R_{ab}m^an^b\,,\\ \bar\Phi_{ij} &= \Phi_{ji}\,.\end{aligned} \tag{25.40b}$$

We are allowed to make transformations of the null tetrad while preserving their inner products, thereby leading to a representation of the proper Lorentz group. The allowed transformations are parameterized by two real parameters (A, ψ) and two complex numbers (a, b) (a total of six real parameters in all, as expected):

i) Boosts: $\ell \to A\ell\,, \quad n \to A^{-1}n\,, \quad m \to m$
ii) Spin rotations in the $m - \bar{m}$ plane: $m \to e^{i\psi}m$, $\ell \to \ell$, $n \to n$
iii) Null rotations around ℓ: $\ell \to \ell$, $m \to m + a\ell$, $n \to n + \bar{a}m + a\bar{m} + |a|^2\ell$
iv) Null rotations around n: $n \to n$, $m \to m + b\ell$, $\ell \to \ell + \bar{b}m + b\bar{m} + |b|^2 n$

The transformations of the spin coefficients and curvature components under these transformations are not difficult to work out. Again, we refer to [25.2, 5, 6] for a more complete discussion.

The relation between the spin coefficients and the curvature components leads to the so-called Newman–Penrose field equations which are a set of 16 complex first-order differential equations. The Bianchi identities, $\nabla_{[a}R_{bc]de} = 0$, are written explicitly as eight complex equations involving both the Weyl and Ricci tensor components, and three real equations involving only Ricci tensor components. See [25.2, 5, 6] for the full set of field equations and Bianchi identities.

25.4.2 The Kerr Spacetime in the Newman–Penrose Formalism

As the prototypical example for an isolated horizon, we now describe the structure of the Kerr black hole horizon. This will also illustrate the utility of the Newman–Penrose formalism when dealing with null surfaces. A detailed study of the various intricate properties of the Kerr spacetime can be found in [25.2]. Here we shall be brief and focus on the essential properties of the horizon.

The Kerr metric with mass M and spin a is usually presented in textbooks as (this is however not the form that Kerr originally derived it)

$$\begin{aligned} \mathrm{d}s^2 = &-\left(1 - \frac{2Mr}{\rho^2}\right)\mathrm{d}t^2 + \frac{\rho^2}{\Delta}\mathrm{d}r^2 \\ &- \frac{4aMr\sin^2\theta}{\rho^2}\mathrm{d}t\,\mathrm{d}\phi + \rho^2\mathrm{d}\theta^2 \\ &+ \frac{\Sigma^2\sin^2\theta}{\rho^2}\mathrm{d}\phi^2\,, \end{aligned} \tag{25.41}$$

where

$$\begin{aligned} \rho^2 &= r^2 + a^2\cos^2\theta\,, \\ \Delta &= r^2 - 2Mr + a^2\,, \\ \Sigma^2 &= (r^2 + a^2)\rho^2 + 2a^2Mr\sin^2\theta\,. \end{aligned} \tag{25.42}$$

This metric has two Killing vectors: a timelike one $\xi^a = (\partial_v)^a$, and a spacelike rotational one $\varphi^a = (\partial_\phi)^a$. Based on the behavior of the metric at large distances, one can assign a mass $M_\infty = M$ and angular momentum $J_\infty = aM$ to the spacetime. There are multiple ways to justify this. Because of the existence of the two Killing vectors, the clearest definition is through the so-called Komar integrals [25.31] based on the two Killing vectors (see also [25.3]). Moreover, again based on the behavior of the gravitational field at infinity, one can assign two sets of higher multipole moments M_k and J_k [25.32–34] which turn out to be fully determined by M and a: $M_k + iJ_k = M(ia)^k$, $k = 2, 3, \ldots$.

As in the original Schwarzschild metric, there are coordinate singularities when $\Delta = 0$. This happens when

$$r = r_\pm = M \pm \sqrt{M^2 - a^2}\,. \tag{25.43}$$

These can be removed by a coordinate transformation $(t, r, \theta, \phi) \to (v, r, \theta, \varphi)$

$$\mathrm{d}v = \mathrm{d}t + \frac{r^2 + a^2}{\Delta}\mathrm{d}r\,, \quad \mathrm{d}\varphi = \mathrm{d}\phi - \frac{a}{\Delta}\mathrm{d}r\,. \tag{25.44}$$

This yields the metric in (v, r, θ, φ) candidate

$$\begin{aligned} \mathrm{d}s^2 = &-\left(1 - \frac{2Mr}{\rho^2}\right)\mathrm{d}v^2 + 2\mathrm{d}v\,\mathrm{d}r - 2a\sin^2\theta\,\mathrm{d}r\,\mathrm{d}\varphi \\ &- \frac{4aMr\sin^2\theta}{\rho^2}\mathrm{d}v\,\mathrm{d}\varphi \\ &+ \rho^2\mathrm{d}\theta^2 + \frac{\Sigma^2\sin^2\theta}{\rho^2}\mathrm{d}\varphi^2\,. \end{aligned} \tag{25.45}$$

The horizon is the three-dimensional surface $r = r_+$ which we shall denote by Δ. The intrinsic metric q_{ab} on Δ in (v, θ, φ) coordinates is obtained by setting $r = r_+$ and dropping the $\mathrm{d}r$ terms in this metric. Rearranging terms we get

$$\begin{aligned} q_{ab} = &\frac{a^2\sin^2\theta}{\rho_+^2}\left(\nabla_a v - \Omega^{-1}\mathrm{d}\varphi\right)\left(\nabla_b v - \Omega^{-1}\mathrm{d}\varphi\right) \\ &+ \rho_+^2\nabla_a\theta\nabla_b\theta\,, \end{aligned} \tag{25.46}$$

Part D | 25.4

where $\rho_+^2 := r_+^2 + a^2\cos^2\theta$ and $\Omega = a/2Mr_+ = a/(r_+^2 + a^2)$. It is easy to verify that this metric has signature $(0++)$ with the degenerate direction being

$$\ell^a\nabla_a = \frac{\partial}{\partial v} + \Omega\frac{\partial}{\partial\varphi}\,. \tag{25.47}$$

Thus, the null normal to Δ acquires an angular velocity term in the presence of spin.

The cross-sections of this manifold, i. e., the surfaces of constant v, are spheres with a Riemannian metric $\widetilde{q}_{ab}$. The area of such a sphere is time independent: $a_\Delta = 4\pi(r_+^2 + a^2)$. It is easy to verify that if we choose a different coordinate v' with the cross-sections still being complete spacelike spheres, the area of each cross-section is still a_Δ.

A suitable choice of the ingoing and outgoing future directed null vectors are

$$\begin{aligned} n^a\nabla_a &= -\left(\frac{r^2+a^2}{\rho^2}\right)\frac{\partial}{\partial r}\,,\\ \ell^a\nabla_a &= \frac{\partial}{\partial v} + \frac{a}{r^2+a^2}\frac{\partial}{\partial\varphi} + \frac{\Delta}{2(r^2+a^2)}\frac{\partial}{\partial r}\,. \end{aligned} \tag{25.48}$$

On $\Delta\ \ell^a$ one agrees with the null direction given in (25.47). The other null vector n^a is clearly null because the metric does not have a $\mathrm{d}r^2$ term, and the scalar factor in n^a is chosen to ensure $\ell\cdot n = -1$. The covariant versions are

$$\begin{aligned} n_a &= \frac{r^2+a^2}{\rho^2}(-\nabla_a v + a\sin^2\theta\nabla_a\varphi)\,,\\ \ell_a &= -\frac{\Delta}{2(r^2+a^2)}\nabla_a v + \frac{\rho^2}{r^2+a^2}\nabla_a r\\ &\quad + \frac{\Delta a\sin^2\theta}{2(r^2+a^2)}\nabla_a\varphi\,. \end{aligned} \tag{25.49}$$

A suitable choice for m^a is

$$\begin{aligned} m_a &= -\frac{a\sin\theta}{\sqrt{2}\widetilde{\rho}}\nabla_a v + \frac{(r^2+a^2)\sin\theta}{\sqrt{2}\widetilde{\rho}}\nabla_a\varphi\\ &\quad + \frac{\mathrm{i}}{\sqrt{2}}\bar{\widetilde{\rho}}\nabla_a\theta\,,\\ m^a\nabla_a &= \frac{a\sin\theta}{\sqrt{2}\widetilde{\rho}}\frac{\partial}{\partial v} + \frac{1}{\sqrt{2}\widetilde{\rho}\sin\theta}\frac{\partial}{\partial\varphi}\\ &\quad + \frac{\mathrm{i}}{\sqrt{2}\widetilde{\rho}}\frac{\partial}{\partial\theta}\,. \end{aligned} \tag{25.50}$$

Here we have defined $\widetilde{\rho} := r + \mathrm{i}a\cos\theta$, so that $\rho^2 = |\widetilde{\rho}|^2$. It is unfortunate that the notation can be confusing. For example the ρ^2 used in the Kerr metric is not to be confused with the spin coefficient ρ. This should hopefully not cause confusion because the spin coefficient ρ will vanish identically, and unless mentioned otherwise, ρ^2 will refer to $r^2 + a^2\sin^2\theta$.

We can now compute the spin coefficients at Δ and for the moment we shall restrict our attention to those spin coefficients which are intrinsic to Δ, i. e., do not require any derivatives transverse to Δ. These are: $\epsilon, \kappa, \pi, \alpha, \beta, \rho, \sigma, \mu, \lambda$. As is typical in tetrad formalisms, we do not need to compute any Christoffel symbols in order to compute any of the spin coefficients; the exterior derivative suffices. Using the definitions of (25.34) we can write the exterior derivatives of ℓ_a, m_a, n_a in terms of the exterior products of the basis vectors. The spin coefficients are then combinations of contractions of the exterior derivatives with the basis vectors. As an example, the acceleration of ℓ^a and its value at the Kerr horizon is

$$\begin{aligned} \epsilon + \bar{\epsilon} &= \ell^b n^a \cdot 2\nabla_{[a}\ell_{b]} = \frac{r_+ - M}{2Mr_+}\\ &= \frac{\sqrt{M^2-a^2}}{2M(M+\sqrt{M^2-a^2})}\,. \end{aligned} \tag{25.51}$$

The other spin coefficients at the horizon turn out to be

$$\begin{aligned} \kappa &= \sigma = \rho = \lambda = \nu = 0\,,\\ \pi &= \alpha + \bar{\beta}\\ &= -\frac{\sqrt{2}ar\sin\theta}{\rho^2\bar{\widetilde{\rho}}}\,. \end{aligned} \tag{25.52}$$

Some of these can be understood on general grounds. First, since ℓ^a is tangent to a smooth surface, it must be hyper-surface orthogonal which means that we must have $\mathrm{Im}\rho = 0$. Furthermore, if a null vector is hypersurface orthogonal, it can be shown to be tangent to a geodesic. This implies $\kappa = 0$. The important condition on physical grounds is that the expansion of ℓ^a vanishes: $\mathrm{Re}\rho = 0$. Thus, as we saw for a Schwarzschild black hole, the cross-sections of Δ are marginally outer trapped surfaces. Now let us turn to the Weyl tensor. It can be shown that the only nonzero component is

$$\Psi_2 = -\frac{M}{(r - \mathrm{i}a\cos\theta)^3}\,. \tag{25.53}$$

Can we now extract from these results general properties of a null surface which should behave like a black

hole horizon in equilibrium? Can we assign physical quantities such as surface gravity, mass, angular momentum, and higher multipole moments? We shall answer these questions in the next subsection. Regarding angular momentum, we note that the spin coefficient which vanishes for $a = 0$ is π. In fact, from the values of the spin coefficient, we can check that for any X^a tangent to the horizon, $X^a \nabla_a \ell^b = X^a \omega_a \ell^b$ where

$$\omega_a := -(\epsilon + \bar{\epsilon}) n_a + \pi m_a + \bar{\pi} \bar{m}_a \,. \tag{25.54}$$

We shall see that the angular part of ω_a yields the angular momentum of the horizon.

25.4.3 A Primer on Null Hyper-Surfaces

The fundamental geometric objects in the theory of isolated horizons are null surfaces. Let us therefore start with with a discussion of the geometry of null surfaces in a Lorentzian manifold. The horizon of a Kerr black hole is a special kind of null surface, namely an expansion and shear-free null surface. Such null surfaces are rather special from a geometrical point of view as we shall now explain.

Consider a smooth submanifold $\widetilde{\mathcal{M}}$ of a spacetime $(\mathcal{M}, g_{ab})$. As earlier in Sect. 25.3.2, we define the first fundamental tensor of $\widetilde{\mathcal{M}}$, i. e., the induced metric h_{ab} as the restriction of g_{ab} to $\widetilde{\mathcal{M}}$: $h_{ab}X^aY^b := g_{ab}X^aY^b$ for any two arbitrary vector fields X^a and Y^b tangent to $\widetilde{\mathcal{M}}$. The submanifold $\widetilde{\mathcal{M}}$ is said to be null when h_{ab} is degenerate. In the nondegenerate case, the ambient covariant derivative operator ∇ induces a natural derivative operator $\mathcal{D}$ on $\widetilde{\mathcal{M}}$. Furthermore, $\mathcal{D}$ is the unique derivative operator compatible with h_{ab}, i. e., $\mathcal{D}_a h_{bc} = 0$.

Can we repeat the steps described in Sect. 25.3.2 for defining the fundamental forms and intrinsic connection on a null surface? When $\widetilde{\mathcal{M}}$ is a null hyper-surface, we shall call ℓ^a a null normal if it is along the degenerate direction of h_{ab}, so that $h_{ab}\ell^a = 0$. We thus encounter a problem in the very first step, i. e., in the decomposition of (25.24). The null normal ℓ^a is also tangent to $\widetilde{\mathcal{M}}$ so that $T_p\widetilde{\mathcal{M}} \cap T_p^{\perp}\widetilde{\mathcal{M}} \neq \{0\}$. The other route to defining $\mathcal{D}$, namely finding the unique derivative operator compatible with h_{ab}, does not work either because a degenerate metric does not uniquely determine a derivative operator. We can still define an intrinsic derivative operator $\mathcal{D}$ if we pick a particular subspace of T_pM transverse to $\widetilde{\mathcal{M}}$. Following [25.35], we first pick a spacelike subspace S_p of $T_p\widetilde{\mathcal{M}}$; for the Kerr horizon, a natural choice would be vectors tangent to the spherical cross-sections of fixed v. There will be two one-dimensional subspaces of null vectors orthogonal to S_p. One of them is $N_p\widetilde{\mathcal{M}}$, the null direction of h_{ab}, and the other will be transverse to $\mathcal{M}$ which we shall call $N_p'\widetilde{\mathcal{M}}$, and we can decompose $T_p\mathcal{M}$ as $T_p\widetilde{\mathcal{M}} \oplus N_p\widetilde{\mathcal{M}} \oplus N_p'\widetilde{\mathcal{M}}$. Associated with a particular null normal ℓ^a we shall pick a vector n^a in the transverse direction by requiring that $\ell \cdot n = -1$. We can then decompose ξ as

$$\xi = \xi^{\top} + \xi^{\perp} \quad \text{where} \quad \begin{cases} \xi^{\top} := \widetilde{\xi} + \alpha\ell \\ \xi^{\perp} := \beta n \end{cases} . \tag{25.55}$$

Here α and β are scalars, $\widetilde{\xi}$ is in S_p at each p. We can then define the intrinsic derivative operator as before: $\mathcal{D}_X Y = (\nabla_X Y)^{\top}$. However, $\mathcal{D}$ would depend on our choice of S_p, and unlike in the nondegenerate case, there is in general no natural canonical choice. There is however one case when this is not an issue, namely when the second fundamental form vanishes, $\nabla_X Y$ is always tangential to $\widetilde{\mathcal{M}}$ and there is no need to decompose $\nabla_X Y$. This happens when

$$\ell_a X^b \nabla_b Y^a = -X^a Y^b \nabla_a \ell_b = 0 \,, \tag{25.56}$$

for any X^a, Y^a tangent to $\widetilde{\mathcal{M}}$. Thus, ℓ_a is covariantly constant on the null surface.

An alternative way to state the same result (emphasized in e.g., [25.36]) is: if we start with a vector tangent to Δ and parallel transport it using the spacetime derivative operator ∇ along a curve lying on Δ, then the vector remains tangent to Δ. Since the parallel transport of Y^a along X^a is defined by $X^a \nabla_a Y^b = 0$, this alternative criterion is also equivalent to $\ell_a X^b \nabla_b Y^a = 0$. We shall see that (25.56) is satisfied for the various kinds of horizons that we shall now define.

25.4.4 Nonexpanding, Weakly Isolated and Isolated Horizons

Having understood null surfaces, we are now ready to define different kinds of isolated horizons with increasingly stronger conditions. We shall start with the minimum set of conditions, namely a marginally trapped tube which is null, and no condition on the ingoing expansion. A smooth three-dimensional null surface Δ is said to be a *nonexpanding horizon* if:

- Δ has topology $S^2 \times \mathbb{R}$, and if $\varpi : S^2 \times \mathbb{R} \to S^2$ is the natural projection, then $\varpi^{-1}(x)$ for any $x \in S^2$ are null curves on Δ.
- The expansion $\Theta_{(\ell)} := q^{ab} \nabla_a \ell_b$ of any null normal ℓ^a of Δ vanishes.

- The Einstein field equations hold at Δ, and the matter stress-energy tensor T_{ab} is such that for any future directed null normal ℓ^a, $-T^a_b\ell^b$ is future causal.

We shall consider only null tetrads adapted to Δ such that, at the horizon, ℓ^a coincides with a null normal to Δ. We shall also consider a foliation of the horizon by spacelike spheres S_v with v a coordinate on the horizon which is also an affine parameter along ℓ: $\mathcal{L}_\ell v = 1$; S shall denote a generic spherical cross-section of Δ. Null rotations about ℓ^a correspond to changing the foliation.

This deceptively simple definition of a nonexpanding horizon leads to a number of important results which we state here without proof, most of which are however well illustrated by the Kerr example discussed earlier:

1. Any null normal ℓ^a is a symmetry of the intrinsic degenerate metric q_{ab} on Δ: $\mathcal{L}_\ell q_{ab} = 0$.
2. The null normal of Δ is only given to be expansion free. However, a nonexpanding horizon is also shear free. To show this, we use the Raychaudhuri

$$\mathcal{L}_\ell \Theta_\ell = \kappa_\ell \Theta_{(\ell)} - \tfrac{1}{2}\Theta^2_{(\ell)} - |\sigma|^2 - R_{ab}\ell^a\ell^b \,. \tag{25.57}$$

Setting $\Theta_{(\ell)} = 0$, and observing that the energy condition implies $R_{ab}\ell^a\ell^b \geq 0$, we get that the sum of two nonnegative quantities must vanish.

$$|\sigma|^2 + R_{ab}\ell^a\ell^b = 0 \,. \tag{25.58}$$

This can only happen if $\sigma = 0$ and $R_{ab}\ell^a\ell^b = 0$ on the horizon. Thus, the full projection of $\nabla_a\ell_b$ on Δ vanishes, and as we saw in Sect. 25.4.3, this is just the condition required to ensure that the induced derivative operator on Δ is well defined.
3. The Weyl tensor components Ψ_0 and Ψ_1 vanish on the horizon. This implies that Ψ_2 is an invariant on Δ as long as the null tetrad is adapted to the horizon; it is automatically invariant under boosts and spin rotations (it has spin weight 0), and it is invariant under null rotations around ℓ because Ψ_0 and Ψ_1 vanish. Similarly, the Maxwell field component ϕ_0 vanishes on the horizon, and ϕ_1 is invariant on Δ. Both Ψ_2 and ϕ_1 are also time independent on the horizon.
4. There exists a 1-form ω_a such that, for any vector field X^a tangent to Δ,

$$X^a\nabla_a\ell^b = X^a\omega_a\ell^b \,. \tag{25.59}$$

The 1-form ω_a plays a fundamental role in what follows. The pullback of ω_a to the cross-sections S will be denoted by $\tilde{\omega}_a$.
5. The surface gravity of ℓ is

$$\tilde{\kappa}_{(\ell)} = \ell^a\omega_a \,. \tag{25.60}$$

We will say that Δ is extremal if $\tilde{\kappa}_{(\ell)} = 0$ and nonextremal otherwise. Here we shall always assume that Δ is nonextremal. The curl and divergence of ω carry important physical information. The curl is related to the imaginary part of the Weyl tensor on the horizon

$$\mathrm{d}\omega = \mathrm{Im}\,[\Psi_2]\,{}^2\epsilon \,, \tag{25.61}$$

and its divergence specifies the foliation of Δ by spheres [25.37].
6. By the geometry of Δ, we shall mean the pair $(q_{ab}, \mathcal{D}_a)$. Clearly, q_{ab} yields a Riemannian metric $\widetilde{q}_{ab}$ on the cross-sections of Δ. In turn, $\mathcal{D}_a$ is determined by ω_a and by the unique derivative operator $\widetilde{\mathcal{D}}_a$ compatible with $\widetilde{q}_{ab}$.

We need to strengthen the conditions of a nonexpanding horizon for various physical situations. The minimum extra condition required for black hole thermodynamics and to have a well-defined action principle with Δ as an inner boundary of a portion of spacetime, is formulated as a weakly isolated horizon [25.38]: A weakly isolated horizon $(\Delta, [\ell])$ is a nonexpanding horizon equipped with an equivalence class of null normals $[\ell]$ related by constant positive rescalings and such that

$$\mathcal{L}_\ell\omega_a = 0 \,. \tag{25.62}$$

If we rescale $\ell \to f\ell$, ω_a transforms as $\omega_a \to \omega_a + \partial_a \ln f$. It is thus invariant under constant rescalings and there is a unique ω_a corresponding to the equivalence class $[\ell]$.

The zeroth law holds on weakly isolated horizons, i. e., $\tilde{\kappa}_{(\ell)} = \ell^a\omega_a$ is constant on Δ

$$\mathcal{L}_\ell\omega_a = \ell^b 2\mathcal{D}_{[b}\omega_{a]} + \mathcal{D}_a(\ell^b\omega_b) = \mathcal{D}_a\tilde{\kappa}_{(\ell)} \,. \tag{25.63}$$

In the second step we have used (25.61) to conclude that $\ell^b \times 2\mathcal{D}_{[b}\omega_{a]} = \mathrm{Im}\,\ell^b\,[\Psi_2]\,{}^2\epsilon_{ba} = 0$. Note that under a rescaling $\ell^a \to f\ell^a$, ω_a transforms as $\omega_a \to \omega_a + D_a \ln f$ so that it is invariant under constant rescalings.

Any nonexpanding horizon can be made into a weakly isolated horizon by suitably scaling the null generators. Thus, the restriction to weakly isolated horizons is not a genuine physical restriction. One could go ahead and impose further physical restrictions on the intrinsic horizon geometry by requiring that not only ω_a, but also the full connection $\mathcal{D}_a$ on Δ is preserved by ℓ^a: An isolated horizon $(\Delta, [\ell])$ is thus a nonexpanding horizon equipped with a equivalence class of null normals related by constant positive rescalings such that $[\mathcal{L}_\ell, \mathcal{D}] = 0$ [25.37].

It can be shown that the gauge-invariant geometry of an axisymmetric isolated horizon can be fully specified by the area a_Δ and two sets of multipole moments M_n, J_n for $n = 0, 1, 2, \ldots$ [25.39]. In contrast to the field multipole moments defined at infinity, M_n and J_n are the source multipole moments. We shall not discuss these moments here in any detail, except to say that in the vacuum case, the moments are essentially obtained by decomposing Ψ_2 at Δ into spherical harmonics based on preferred coordinates adapted to the axial symmetry. As at infinity, the Kerr horizon corresponds to a specific choice of these multipoles, but due to the nonlinearity of general relativity, the field and source moments will not generally agree. See, e.g., [25.40–43] for some applications of these multipole moments.

25.4.5 The Near Horizon Geometry

In a number of astrophysical applications where black holes play a role, it is not directly the horizon which is involved but rather the spacetime in the vicinity of the black hole. A especially interesting example of relevance to gravitational wave observations is the case of a binary system consisting of a stellar mass black hole or star orbiting around a much larger black hole (see, e.g., [25.44] for a review of the astrophysics of such systems). As the small particle orbits the large black hole, it effectively *maps* the spacetime in the vicinity of the large black hole, much as the motion of satellites around Earth enables us to map Earth's gravitational potential and therefore its shape. If carried out to sufficient precision, a measurement of gravitational waves from such a system would enable us to determine the gravitational field in the vicinity of the large black hole (see, e.g., [25.45, 46]). One expects that the black hole is well approximated by a Kerr spacetime and we can seek to measure deviations from it thereby testing an important prediction of general relativity. Typical studies on this subject assume the black hole to be Kerr, which is entirely reasonable to a very good approximation (nearby stars or other matter fields might distort the black hole somewhat, but this is expected to be a small effect). However, for mathematical purposes, we could pose the question from a different viewpoint: If we specify the intrinsic geometry of the horizon, then to what extent is the near horizon spacetime determined by the horizon geometry? If we assume the large black hole to be in equilibrium (an excellent approximation for the case we have just described), then it seems reasonable to model the black hole as an isolated horizon. So the question then is: Can we find solutions to the Einstein field equations which admit a generic isolated horizon as an inner boundary? If so, then what is the extra data (beyond the intrinsic horizon geometry) that needs to be specified? (A full solution to the problem would require us to go beyond isolated horizons and to consider small deviations from equilibrium, but we shall not discuss this generalization here.)

It turns out that these questions can be clearly answered if we use the characteristic initial value formulation of Einstein's equations where free data is specified on a set of intersecting null hyper-surfaces [25.6, 47–49]. Consider N dependent variables ψ_I $(I = 1, \ldots, N)$ on a spacetime manifold with coordinates x^a. We shall be concerned with hyperbolic first-order quasi-linear equations of the form

$$\sum_{J=1}^{N} A^a_{IJ}(x, \psi)\partial_a \psi_J + F_I(x, \psi) = 0 \,. \tag{25.64}$$

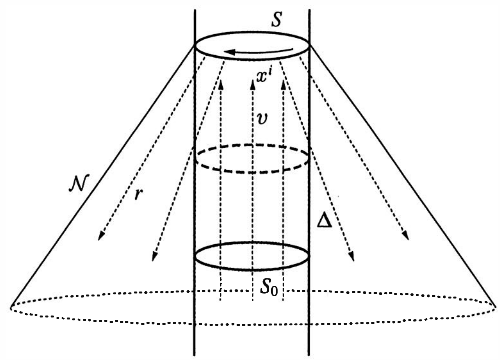


Fig. 25.11 The near horizon coordinates. The isolated horizon is Δ and the transverse null surface is $\mathcal{N}$. The affine parameter along the outgoing null geodesics on $\mathcal{N}$ is r, and v is a coordinate along the null generators on Δ, and x^1 are coordinates on the cross-sections of Δ

In the standard Cauchy problem, one specifies the ψ_I at some initial time. A solution is then guaranteed to be unique and to exist at least locally in time. The characteristic formulation considers a pair of null surfaces $\mathcal{N}_0$ and $\mathcal{N}_1$ whose intersection is a codimension-2 spacelike surface S. It turns out to be possible to specify appropriate data on the null surfaces and on S such that the above system of equations is well posed and has a unique solution, at least locally near S.

In our case, the appropriate free data is specified on the horizon and on an outgoing past light cone originating from a cross-section of the horizon. Such a construction in the context of isolated horizons was first studied by *Lewandowski* [25.50] who characterized the general solution of Einstein equations admitting an isolated horizon. This was worked out in detail in [25.51] which we follow here; similar results from a somewhat different perspective are discussed in [25.52]. The general scenario is sketched in Fig. 25.11. We consider a portion of the horizon Δ which is isolated, in the sense that no matter and/or radiation is falling into this portion of the horizon. For a cross-section S, the past-outgoing light cone is denoted by $\mathcal{N}$. The null generators of Δ and $\mathcal{N}$ are parameterized by v and r, respectively; x^i are coordinates on S. This leads to a coordinate system (v, r, x^i) which is valid till the null geodesics on $\mathcal{N}$ start to cross. The field equations are solved in a power series in r away from the horizon.

Let us now assume that the vacuum Einstein equations hold in a neighborhood of the horizon Δ. Following [25.53], we introduce a coordinate system and null tetrad in the vicinity of Δ analogous to the Bondi coordinates near null infinity. See Fig. 25.11. Choose a particular null normal ℓ^a on Δ. Let v be the affine parameter along ℓ^a so that $\ell^a\nabla_a v = 1$. Let S_v denote the spheres of constant v. Introduce coordinates x^i ($i = 2, 3$) on any one S_v (call this sphere S_0) and require them to be constant along ℓ^a: $\ell^a\nabla_a x^i = 0$; this leads to a coordinate system (v, x^i) on Δ. Let n^a be a future directed inward pointing null vector orthogonal to the S_v and normalized such that $\ell \cdot n = -1$. Extend n^a off Δ geodesically, with r being an affine parameter along $-n^a$; set $r = 0$ at Δ. This yields a family of null surfaces $\mathcal{N}_v$ parameterized by v and orthogonal to the spheres S_v. Set (v, x^i) to be constant along the integral curves of n^a to obtain a coordinate system (v, r, x^i) in a neighborhood of Δ. Choose a complex null vector m^a tangent to S_0. Lie drag m^a along ℓ^a

$$\mathcal{L}_\ell m^a = 0 \quad \text{on } \Delta\,. \tag{25.65}$$

We thus obtain a null tetrad $(\ell, n, m, \bar{m})$ on Δ. Finally, parallel transport ℓ and m along $-n^a$ to obtain a null tetrad in the neighborhood of Δ. This construction is fixed up to the choice of the x^i and m^a on an initial cross-section S_0. We are allowed to perform an arbitrary spin transformation $m \to e^{i\psi} m$ on S_0.

With the Bondi-like coordinate system in hand, we can now in principle use the coordinate basis vectors in the (v, r, x^i) coordinates to construct an arbitrary null tetrad near the horizon. The evolution equations for the component functions of the null tetrad will follow from the above construction. Let us start with n_a and n^a. We have the family of null surfaces $\mathcal{N}_v$ parameterized by v; n_a is normal to the $\mathcal{N}_v$, and r is an affine parameter along $-n^a$. This implies that we can choose

$$n_a = -\partial_a v \quad \text{and} \quad n^a\nabla_a := \Delta = -\frac{\partial}{\partial r}\,. \tag{25.66}$$

To satisfy the inner-product relations $\ell^a n_a = -1$ and $m^a n_a = 0$, the other basis vectors must be of the form

$$\ell^a\nabla_a := D = \frac{\partial}{\partial v} + U\frac{\partial}{\partial r} + X^i\frac{\partial}{\partial x^i}\,,$$
$$m^a\nabla_a := \delta = \Omega\frac{\partial}{\partial r} + \xi^i\frac{\partial}{\partial x^i}\,. \tag{25.67}$$

The frame functions U, X^i are real while Ω, ξ^i are complex. We wish to now specialize to the case when ℓ^a is a null normal of Δ so that the null tetrad is adapted to the horizon. Since ∂_v is tangent to the null generators of Δ, this clearly requires that U, X^i must vanish on the horizon. Similarly, we want m^a to be tangent to the spheres S_v at the horizon, so Ω should also vanish on Δ. Thus, U, X^i, Ω are all $\mathcal{O}(r)$ functions.

To expand the metric in powers of r, we start with the frame fields, and the radial frame equations derived from the commutation relations. The strategy is the same as for the spin coefficients. The radial equations give us the first radial derivatives by substituting the horizon values on the right-hand side, and taking higher derivatives leads to the higher order terms. The calculations are straightforward and lead to the following expansions

$$U = r\tilde{\kappa} + r^2\left(2\left|\pi^{(0)}\right|^2 + \mathrm{Re}\left[\Psi_2^{(0)}\right]\right) + \mathcal{O}(r^3)\,, \tag{25.68a}$$

$$\Omega = r\bar{\pi}^{(0)} + r^2\left(\mu^{(0)}\bar{\pi}^{(0)} + \bar{\lambda}^{(0)}\pi^{(0)} + \tfrac{1}{2}\bar{\Psi}_3^{(0)}\right) + \mathcal{O}(r^3)\,, \tag{25.68b}$$

$$X^i = \bar{\Omega}\xi^i_{(0)} + \Omega\bar{\xi}^i_{(0)} + \mathcal{O}(r^3)\,, \tag{25.68c}$$

$$\xi^i = \left[1 + r\mu^{(0)} + r^2\left((\mu^{(0)})^2 + |\lambda^{(0)}|^2\right)\right]\xi^i_{(0)} + \left[r\bar{\lambda}^{(0)} + r^2\left(2\mu^{(0)}\bar{\lambda}^{(0)} + \tfrac{1}{2}\bar{\Psi}_4^{(0)}\right)\right]\bar{\xi}^i_{(0)} + \mathcal{O}(r^3)\,. \tag{25.68d}$$

The contravariant metric is seen to be given in terms of the frame fields as follows:

$$g^{rr} = 2(U + |\Omega|^2)\,, \quad g^{vr} = 1\,, \tag{25.69a}$$

$$g^{ri} = X^i + \bar{\Omega}\xi^i + \Omega\bar{\xi}^i\,, \quad g^{ij} = \xi^i\bar{\xi}^j + \bar{\xi}^i\xi^j\,. \tag{25.69b}$$

The null cotetrad can be calculated easily up to $\mathcal{O}(r^2)$

$$n = -\mathrm{d}v\,, \tag{25.70}$$

$$\ell = \mathrm{d}r - \left(\tilde{\kappa}r + \mathrm{Re}\left[\Psi_2^{(0)}\right]r^2\right)\mathrm{d}v - \left(\pi^{(0)}r + \tfrac{1}{2}\Psi_3^{(0)}r^2\right)\xi_i^{(0)}\,\mathrm{d}x^i - \left(\bar{\pi}^{(0)}r + \tfrac{1}{2}\bar{\Psi}_3^{(0)}r^2\right)\bar{\xi}_i^{(0)}\,\mathrm{d}x^i\,, \tag{25.71}$$

$$m = -\left(\bar{\pi}^{(0)}r + \tfrac{1}{2}\Psi_3^{(0)}r^2\right)\mathrm{d}v + (1 - \mu^{(0)}r)\xi_i^{(0)}\,\mathrm{d}x^i - \left(\bar{\lambda}^{(0)}r + \tfrac{1}{2}\bar{\Psi}_4^{(0)}r^2\right)\bar{\xi}_i^{(0)}\,\mathrm{d}x^i\,. \tag{25.72}$$

Here $\xi_i^{(0)}$ are defined by the relations $\xi_i^{(0)}\xi^i_{(0)} = 0$ and $\xi_i^{(0)}\bar{\xi}^i_{(0)} = 1$; it will be convenient to set $m_a^{(0)} := \xi_i^{(0)}\partial_a x^i$. In powers of r, the metric is

$$g_{ab} = -2\ell_{(a}n_{b)} + 2m_{(a}\bar{m}_{b)} = g_{ab}^{(0)} + g_{ab}^{(1)}r + \tfrac{1}{2}g_{ab}^{(2)}r^2 + \cdots\,, \tag{25.73}$$

where

$$g_{ab}^{(0)} = 2\partial_{(a}r\partial_{b)}v + 2m_{(a}^{(0)}\bar{m}_{b)}^{(0)}\,, \tag{25.74}$$

$$g_{ab}^{(1)} = -\left(2\tilde{\kappa}\partial_{(a}v\partial_{b)}v + 4\pi^{(0)}m_{(a}^{(0)}\partial_{b)}v + 4\bar{\pi}^{(0)}\bar{m}_{(a}^{(0)}\partial_{b)}v + 4\mu^{(0)}m_{(a}^{(0)}\bar{m}_{b)}^{(0)} + 2\lambda^{(0)}m_{(a}^{(0)}m_{b)}^{(0)} + 2\bar{\lambda}^{(0)}\bar{m}_{(a}^{(0)}\bar{m}_{b)}^{(0)}\right), \tag{25.75}$$

$$\begin{aligned} g_{ab}^{(2)} &= 4\left(|\pi^{(0)}|^2 - \mathrm{Re}\left[\Psi_2^{(0)}\right]\right)\partial_{(a}v\partial_{b)}v \\ &+ 4\left(\mu^{(0)}\pi^{(0)} + \lambda^{(0)}\bar{\pi}^{(0)} - \Psi_3^{(0)}\right)m_{(a}^{(0)}\partial_{b)}v \\ &+ 4\left(\mu^{(0)}\bar{\pi}^{(0)} + \bar{\lambda}^{(0)}\pi^{(0)} - \bar{\Psi}_3^{(0)}\right)\bar{m}_{(a}^{(0)}\partial_{b)}v \\ &+ 4\left((\mu^{(0)})^2 + \left|\lambda^{(0)}\right|^2\right)m_{(a}^{(0)}\bar{m}_{b)}^{(0)} \\ &+ \left(4\mu^{(0)}\lambda^{(0)} - 2\Psi_4^{(0)}\right)m_{(a}^{(0)}m_{b)}^{(0)} \\ &+ \left(4\mu^{(0)}\bar{\lambda}^{(0)} - 2\bar{\Psi}_4^{(0)}\right)\bar{m}_{(a}^{(0)}\bar{m}_{b)}^{(0)}\,. \end{aligned} \tag{25.76}$$

Iterating this procedure to higher orders is, in principle, straightforward. This calculation provides a starting point for a number of ongoing works in applying isolated and quasi-local horizons to astrophysical situations.

25.4.6 Angular Momentum, Mass, and the First Law for Isolated Horizons

The first law for black holes, and black hole thermodynamics in general, was developed by *Bekenstein* [25.54], and by *Bardeen* et al. [25.55] in 1973 in analogy with the laws of thermodynamics. The zeroth law says that the surface gravity is constant over the black hole horizon. We have already seen that this is true for a weakly isolated horizon: the surface gravity $\kappa_\ell = \ell^a\omega_a$ is constant on Δ. The main difference with the standard formulation for a globally stationary surface gravity refers to the globally defined stationary Killing vector which is normalized to have unit norm at infinity. A weakly isolated horizon does not refer to any globally defined Killing vector. The standard formulation of the second law says that the area of the event horizon can never decrease in time. The area of a nonexpanding horizon is constant, thus the second law is trivial in this context.

Let us now turn to the first law. The standard formulation for a stationary black hole is

$$\delta M = \frac{\kappa}{8\pi}\delta a + \Omega\delta J\,. \tag{25.77}$$

Here M is the mass measured at spatial infinity, κ is the surface gravity at the event horizon but using a vector field normalized at infinity, a is the area of the horizon, Ω is the angular velocity at the horizon, and J is the angular momentum at infinity. Electromagnetic fields can also be included and leads to additional terms. This is

reasonable when applied to a globally stationary spacetime, but clearly needs to be refined for a black hole in equilibrium locally in an otherwise dynamical spacetime. It turns out that it is possible to formulate the first law for isolated horizons using quantities defined only at the horizon, without reference to the behavior of fields at infinity [25.36, 38, 56, 57]. The setup is a variational problem in a portion of spacetime bounded inside by an axisymmetric weakly isolated horizon $(\Delta, [\ell^a])$ (the extension to multiple horizons is straightforward).

Before talking about the first law, we need to first have suitable notions of the horizon angular momentum and mass. We begin with angular momentum. Fix an axial symmetry φ^a at the horizon. This means that φ^a must preserve (i) the equivalence class $[\ell^a]$ of null normals that is prescribed for the weakly isolated horizon, (ii) the intrinsic metric q_{ab}, and (iii) the 1-form ω_a. Furthermore, φ^a should commute with ℓ^a and it should look like a rotational vector field in that it should have closed integral curves, an affine length of 2π, and should vanish at exactly two null horizon generators. Consider then a rotational vector field ϕ^a in spacetime such that at Δ it is equal to the fixed symmetry: $\phi^a|_\Delta = \varphi^a$. At infinity, we require that it approach some fixed rotational symmetry of the asymptotic flat metric. We then need to find the Hamiltonian H_ϕ (Recall that in a phase space, a Hamiltonian is responsible for generating time evolution via the Poisson bracket. Thus, for any function F in a phase space, $\dot{F} = \{H, F\}$. In the present case, the phase space consists of gravitational (and other) fields which satisfy the appropriate boundary conditions.) which generates motions along ϕ^a. The Hamiltonian can be shown to reduce to integrals over the boundaries at the horizon and at infinity. The term at Δ is identified with the angular momentum of Δ

$$J_\phi^\Delta = -\frac{1}{8\pi} \oint_S \varphi^a \omega_a {}^2\epsilon \,. \tag{25.78}$$

Similarly, the notion of energy corresponds to time translations. Thus, we consider time evolution vector fields t^a on spacetime such that at the horizon, it approaches a general symmetry vector $A\ell^a + \Omega\varphi^a$. Here the coefficients A and Ω are constants on Δ, but are allowed to vary in phase space. The strategy is then again to compute the surface term at Δ in the Hamiltonian H_t which generates motions along t^a, and the surface term at Δ is to be identified as the energy E_t^Δ. Surprisingly, it turns out that motions along t^a are not always Hamiltonian, and in fact, the Hamiltonian exists if and only if

$$\delta E_t^\Delta = \frac{\kappa_t}{8\pi}\delta a_\Delta + \Omega_t \delta J_\Delta \,. \tag{25.79}$$

This is just the first law, and a vector field t^a is said to be *permissible* if the first law for E_t^Δ is satisfied. If the first law is satisfied, it is easy to see that κ_t, Ω_t can depend only on the horizon quantities (a_Δ, J_Δ), and must satisfy integrability condition

$$\frac{\partial \kappa_t(a_\Delta, J_\Delta)}{\partial J_\Delta} = \frac{\partial \Omega(a_\Delta, J_\Delta)}{\partial a_\Delta} \,. \tag{25.80}$$

This ensures that the right-hand side of (25.79) can be integrated to yield an exact variation, and to thus have a well-defined E_t^Δ.

A preferred choice of t^a at the horizon can be obtained by choosing $\kappa(a_\Delta, J_\Delta)$ and $\Omega(a_\Delta, J_\Delta)$ to have the same functional dependence on (a_Δ, J_Δ) as in the Kerr spacetime

$$\kappa = \frac{R_\Delta^4 - 4J_\Delta^2}{2R_\Delta^3\sqrt{R_\Delta^4 + 4J_\Delta^2}} \,, \quad \Omega = \frac{2J_\Delta}{R_\Delta\sqrt{R_\Delta^4 + 4J_\Delta^2}} \,. \tag{25.81}$$

This leads to the horizon mass

$$M_\Delta = \frac{1}{2R_\Delta}\sqrt{R_\Delta^4 + 4J_\Delta^2} \,. \tag{25.82}$$

Finally, we note that J_φ^Δ is independent of the choice of cross-section S, and even though it requires a weakly isolated horizon to carry out the Hamiltonian computation, formula (25.78) itself is well defined even on a nonexpanding horizon. If a cross-section S of Δ is contained within a spatial hyper-surface Σ, and if $\widehat{r}^a$ is the unit spacelike normal to S in Σ and K_{ab} is the extrinsic curvature of Σ, then we can rewrite J_φ^Δ as

$$J_\Delta^\varphi = \frac{1}{8\pi}\oint_S K_{ab}\varphi^a \widehat{r}^b \, {}^2\epsilon \,. \tag{25.83}$$

This is particularly convenient in numerical relativity where one routinely located MTSs on spatial Cauchy surfaces, and would like to use them to characterize the properties of a black hole in real time while the simulation is in progress [25.58]. The computation of angular momentum is a surface integral over the MTS and the mass is just an algebraic expression. These methods are now in common use in numerical simulations.

25.5 Dynamical Horizons

25.5.1 The Area Increase Law

The second law for event horizons states that the area of an event horizon can never decrease in time. This was first suggested by Bekenstein and is the starting point for associating the area of a black hole horizon with entropy. We have thus far not talked about the area increase law for quasi-local horizons except in the context of isolated horizons where it is essentially trivial. For a dynamical horizon, the area increase law follows easily from the fact that $\Theta_{(n)} < 0$. Let $\mathcal{H}$ be a dynamical horizon and S a generic MTS on it. Let r^a be the spacelike unit normal to S within $\mathcal{H}$; this is not to be confused with the unit normal $\widehat{r}^a$ to S which lies on a spatial hyper-surface intersecting $\mathcal{H}$ as in Fig. 25.9. Let τ^a be the unit-timelike vector normal to $\mathcal{H}$. We also assume that r^a points outward, in the sense that if $\widehat{r}^a$ is the outward normal to S on a spatial hyper-surface Σ, then $r^a\widehat{r}_a > 0$. Finally, let $\mathcal{H}$ be bounded by the cross-sections S_1 and S_2. Then, the suitable choices for the out- and ingoing null normals to S are

$$\ell^a = \frac{\tau^a + r^a}{\sqrt{2}}\,, \quad \tau^a = \frac{\tau^a - r^a}{\sqrt{2}}\,. \tag{25.84}$$

Then, if q_{ab} is the intrinsic Riemannian metric on S,

$$\sqrt{2}q^{ab}\nabla_a r_b = q^{ab}\nabla_a(\ell^a - \tau^a) = \Theta_{(\ell)} - \Theta_{(n)} > 0\,. \tag{25.85}$$

Thus, the area element on S and thus its area increases along r^a. (Though we shall not pursue this further, one could imagine relaxing the requirement $\Theta_{(n)} < 0$ by an average on S and still obtain the same result.) This is the area increase law for dynamical horizons.

We can go further and ask whether it is possible to obtain a *physical process* version of the area increase law which relates the increase in area from an initial cross-section S_1 to a later time S_2 to the amount of energy or radiation falling into the black hole between S_1 and S_2. An early result along these lines was proved by *Hartle* and *Hawking* in 1972 [25.59] for perturbations of the event horizon. We note however that such a law does not exist for event horizons in general. A case in point being the Vaidya solution, which as we have seen, grows in flat space when nothing falls into the black hole. Let us then consider the area increase law on a dynamical horizon $\mathcal{H}$. Let us first consider angular momentum and the change in angular momentum from S_1 to S_2.

Since $\mathcal{H}$ is a spacelike hyper-surface we have, as discussed earlier, the induced metric h_{ab}, the associated derivative operator D_a, and the extrinsic curvature K_{ab}. As before, let τ^a be the unit timelike normal to $\mathcal{H}$ and let r^a be the unit spacelike normal to a cross-section S within $\mathcal{H}$. The Einstein equations then show that (h_{ab}, K_{ab}) cannot be specified freely, but must satisfy the Hamiltonian and momentum constraint equations (see, e.g., [25.3]). The momentum constraint is

$$D_b(K^{ab} - Kh^{ab}) = 8\pi T^{bc}n_c h^a_b\,. \tag{25.86}$$

Contracting both sides with a rotational vector field φ^a and integrate by parts to obtain

$$\begin{aligned}&\frac{1}{8\pi}\oint_{S_2} K_{ab}\varphi^a r^b\,\mathrm{d}^2V - \frac{1}{8\pi}\oint_{S_1} K_{ab}\varphi^a r^b\\ &= \int_{\mathcal{H}}\left(T_{ab}\tau^a\varphi^b + \frac{1}{16\pi}P^{ab}\mathcal{L}_\varphi h_{ab}\right),\end{aligned} \tag{25.87}$$

where $P^{ab} = K^{ab} - Kh^{ab}$. We then identify the angular momentum of a cross-section S as

$$J_S^{(\varphi)} = -\frac{1}{8\pi}\int_S K_{ab}\varphi^a r^b\,\mathrm{d}^2V\,. \tag{25.88}$$

Equation (25.87) is thus a *balance* law for angular momentum [25.27, 28]. It relates $J_{S_2}^{(\varphi)} - J_{S_1}^{(\varphi)}$ with the gravitational and matter flux crossing the horizon between S_2 and S_1. We note that if φ^a is a Killing vector of the metric h_{ab}, then the gravitational contribution vanishes identically.

A similar (but more involved calculation) leads to a balance law for the change in area, or rather for the area radius R defined as $R = \sqrt{a/4\pi}$ [25.27, 28]

$$\begin{aligned}\frac{R_2}{2} - \frac{R_1}{2} &= \int_{\mathcal{H}} T_{ab}\tau^a\xi^b\,\mathrm{d}^3V\\ &+ \frac{1}{16\pi}\int_{\mathcal{H}} N_r\left(|\sigma|^2 + 2|\zeta|^2\right)\mathrm{d}^3V\,.\end{aligned} \tag{25.89}$$

Here $|\sigma|^2 := \sigma_{ab}\sigma^{ab}$ with σ_{ab} being the shear of the outgoing null vector $\ell^a = \tau^a + r^a$; $|\zeta|^2 := \zeta_a\zeta^b$ where $\zeta^a = h^{ab}r^c\nabla_c\ell_b$; and $N_r := |D_aRD^aR|^{1/2}$. Equation (25.89) is

the desired area balance law. Again the flux terms consist of a matter and gravitational contributions. The integrand in the gravitational part is manifestly local and nonnegative, and the matter contribution is positive if the dominant energy condition holds. The gravitational contribution in fact vanishes in spherical symmetry as it should. Along similar lines, *Jaramillo* and *Gourgoulhon* obtained a second-order differential equation for the area [25.60] which leads to a causal evolution of the area subject to initial conditions (the analogous evolution of the area of an event horizon [25.61] turns out to be teleological in the sense that it requires one to specify a boundary condition near future timelike infinity).

Let us conclude with few words about angular momentum. We have identified (25.88) with the angular momentum of a cross-section of the dynamical horizon, while earlier we had identified (25.83) with the angular momentum of a cross-section of an isolated horizon. It can be shown that when φ^a is a symmetry of the intrinsic metric, then the two agree. Recently, it was shown by *Jaramillo* et al. [25.62], that for an axisymmetric MOTS which is stably outermost, and if the spacetime satisfies the dominant energy condition, then the angular momentum J defined as above, satisfies the inequality $|J| \leq a/8\pi$ where a is the area of S. Here, axisymmetry is imposed only at S and not globally. This surprising result further validates the identification of (25.88) as the angular momentum.

25.5.2 Uniqueness Results for Dynamical Horizons

In earlier sections, we have already discussed the location and time evolution of trapped and marginally trapped surfaces. We now discuss further results obtained by *Ashtekar* and *Galloway* [25.63] related to the issue of uniqueness of marginally trapped surfaces in the dynamical case.

The first result, proved in [25.63], concerns the uniqueness of the foliation of a dynamical horizon by marginally trapped surfaces. Let $\mathcal{H}$ be a dynamical horizon foliated by a set marginally trapped surfaces S_t with t being a continuous real parameter taking values within an open interval. Let S be a *weakly* trapped surface in $\mathcal{H}$, i.e., both of its expansions are nonpositive: $\Theta_{(\ell)} \leq 0$, $\Theta_{(n)} \leq 0$ (in particular, S could be a marginally trapped surface). Then S must in fact be a marginally trapped surface and must coincide with one of the S_t. This shows that the foliation of $\mathcal{H}$ by the S_t must be unique. As a corollary, consider, as in Fig. 25.9, an MTT generated by MTSs S_t which lie in spacelike surfaces Σ_t. This is a situation common in numerical relativity where one uses Σ_t for time evolution, and one locates MTSs on them. If the intersection $\Sigma' \bigcap \mathcal{H}$ is not one of the S_t, then it cannot be a marginally trapped surface. This is illustrated in Fig. 25.7 for the Vaidya spacetime. The intersection of the nonsymmetric spatial hyper-surface with the spherically symmetric dynamical horizon is the red surface, and it is not a marginally (or weakly) trapped surface. Thus, if we had a different foliation $\Sigma'_{t'}$ and we locate MTSs on these spatial hyper-surfaces, then we would end up with a *different* dynamical horizon $\mathcal{H}'$. We note again the dramatic difference for an isolated horizon where every spherical cross-section is a marginally outer trapped surface.

The question then arises: how different can $\mathcal{H}'$ be from $\mathcal{H}$? A partial answer is provided by the next result [25.63]: *There are no weakly trapped surfaces contained in the past domain of dependence $D^-(\mathcal{H})$ (apart from those which make up $\mathcal{H}$ itself).* (The past domain of dependence $D^-(\mathcal{H})$ is the set of spacetime points p such that every future causal curve from p intersects $\mathcal{H}$.) In particular, this result rules out a dynamical horizon contained in $D^-(\mathcal{H}) - \mathcal{H}$. Once again, this result is illustrated by the Vaidya example discussed earlier in Fig. 25.4. The nonsymmetric surface S in this figure is partly inside and partly outside the spherically symmetric dynamical horizon. More generally, weakly trapped surfaces lying partly inside $D^-(\mathcal{H})$ are not ruled out.

25.6 Outlook

As discussed in Sect. 25.1, ever since the discovery of the Schwarzschild solution almost a century ago, research in black holes has led to seminal developments in theoretical, observational, and computational physics. This includes the singularity theorems, black hole thermodynamics, the uniqueness theorems, the cosmic censorship hypothesis, black hole entropy calculations in various approaches to quantum gravity, the various astrophysical phenomena involving black holes, and the recent results from numerical relativity.

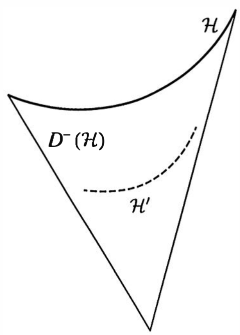


Fig. 25.12 This situation is ruled out by the uniqueness results of Ashtekar and Galloway. If there is a dynamical horizon $\mathcal{H}$, then there cannot be another dynamical horizon $\mathcal{H}'$ lying completely in the past domain of dependence $D^-(\mathcal{H})$

The framework of quasi-local horizons, which takes trapped and marginally trapped surfaces as its starting point, provides a unified approach for studying various aspects of black hole physics. While we have only touched upon a few aspects and applications of this framework, the material presented in this chapter will hopefully motivate the reader to delve further into the subject. An important theme in this discussion is that to base our understanding of black holes on lessons from stationary cases can be misleading. Dynamical situations have some essentially different features and intuition from stationary examples can easily lead us astray. We have illustrated this by the Schwarzschild and Vaidya examples. In this chapter, we have introduced trapped surfaces and various kinds of quasi-local horizons through examples. The eventual goal of these studies (from a physics viewpoint) is to understand the properties of the surface of a black hole. We have discussed the inadequacy of event horizons for this purpose due to its teleological properties, and it is desirable to find a suitable replacement. Penrose's trapped surfaces and the boundary of the trapped region seem ideally suited for this task and lead naturally to the various definitions of quasi-local horizons. The simplest example is of course the Schwarzschild black hole. In this example, the boundary of the trapped region agrees with the event horizon, and both notions give rise to the same physical ideas. Difficulties arise however when we consider nonstationary black holes. We illustrated this through the imploding Vaidya spacetime. The intuitively obvious horizon, the analog of the $r = 2M$ hyper-surface in Schwarzschild is a spherically symmetry dynamical horizon and it is separated from the event horizon. However, the nonspherically symmetric outer trapped surfaces extend up to the event horizon. To make matters more complicated, trapped surfaces do not extend all the way to the event horizon. We used these examples as motivations for general definitions and we reviewed some basic results regarding trapped surfaces and quasi-local horizons. We saw that marginally trapped surfaces are not as ill-behaved as one might think, and under physically reasonable conditions, they do evolve smoothly. The equilibrium case, described by isolated horizons, is also of great interest. It covers a wide variety of situations where a black hole is in equilibrium in a dynamical spacetime and is the best understood quasi-local horizon. Finally, in the dynamical case, we saw that one can assign physical quantities such as mass, angular momentum, and fluxes through dynamical horizons.

There are a number of topics that we have not discussed. In particular, we have not discussed the various applications of these notions in numerical relativity. Similarly, our discussion has been restricted to the classical world and we have not talked about, e.g., the quantization of isolated horizons and the black hole entropy calculations. Even in the discussion of the mathematical properties of quasi-local horizons, we have discussed black hole multipole moments, symmetries, and inclusion of various kinds of matter fields only very briefly. Another significant omission is the discussion of black holes near equilibrium. Reviews of these topics can be found in, e.g. [25.64–66].

We have seen that the outstanding question in this field is the nonuniqueness of dynamical horizons. We have discussed some restrictions on where dynamical horizons can be located. However, it is a fact that there are a multitude of smooth marginally trapped tubes and dynamical horizons in a general black hole spacetime, and there seems to be no obvious way of picking a preferred one. While there could be reasonable choices in specific examples, there does not seem to be any general solution to this problem. While one can get away with using event horizons in globally stationary spacetimes, choosing to live with event horizons in general is not an option because of its global properties. One could perhaps consider dealing with the full set of dynamical horizons and marginally trapped tubes. Many of them are smooth one can study them individually; e.g., the dynamical horizon flux formulae apply to all dynamical horizons equally. However, if we wish to assign physical properties to the black hole such as mass, angular momentum, fluxes, and higher multipoles, which one should we choose? We will generally get differ-

ent results depending on our choice. The boundary of the trapped region could be a reasonable alternative, but as we have seen, it is generally not a marginally trapped tube itself. It is also not clear whether one can use this boundary to study, say, black hole thermodynamics and other physical phenomena that we believe are true for black holes, and besides, this boundary also has a number of global properties and is difficult to locate (thus making it not better than event horizons in many regards). An interesting possibility, suggested by Bengtsson and Senovilla, is the *core* of the black hole region, i. e., the portion of the trapped region where all trapped surfaces penetrate. Removing the core would then completely eliminate all trapped surfaces. If it is indeed a proper subset of the trapped region, then is its boundary a dynamical horizon? There are indications that the region $r \leq 2M(v)$ of Vaidya is such a core, and its boundary is the spherically symmetric dynamical horizon. However, the core is not unique and some cores are not spherically symmetric [25.15]. It is an interesting open question whether this idea can be developed further.

References

25.1 S. Hawking, G. Ellis: *The Large Scale Structure of Space-Time*, Cambridge Monographs on Mathematical Physics (Cambridge Univ. Press, Cambridge 1973)
25.2 S. Chandrasekhar: *The Mathematical Theory of Black Holes*, Oxford Classic Texts in the Physical Sciences (Oxford Univ. Press, Oxford 1985)
25.3 R.M. Wald: *General Relativity* (Univ. Chicago Press, Chicago 1984)
25.4 J. Lee: *Riemannian Manifolds: An Introduction to Curvature*, Graduate Texts in Mathematics (Springer, New York 1997)
25.5 R. Penrose, W. Rindler: *Spinors and Spacetime: 1. Two-Spinor Calculus and Relativistic Fields*, Cambridge Monographs on Mathematical Physics (Cambridge Univ. Press, Cambridge 1984)
25.6 J. Stewart: *Advanced General Relativity*, Cambridge Monographs on Mathematical Physics (Cambridge Univ. Press, Cambridge 1991)
25.7 R.M. Wald, V. Iyer: Trapped surfaces in the Schwarzschild geometry and cosmic censorship, Phys. Rev. D **44**, 3719–3722 (1991)
25.8 P. Vaidya: The gravitational field of a radiating star, Proc. Indian Acad. Sci. A **33**, 264 (1951)
25.9 J. Thornburg: Event and apparent horizon finders for 3 + 1 numerical relativity, Living Rev. Relativ. **10**, 3 (2007)
25.10 E. Schnetter, B. Krishnan: Non-symmetric trapped surfaces in the Schwarzschild and Vaidya spacetimes, Phys. Rev. D **73**, 021502 (2006)
25.11 A.B. Nielsen, M. Jasiulek, B. Krishnan, E. Schnetter: The slicing dependence of non-spherically symmetric quasi-local horizons in Vaidya spacetimes, Phys. Rev. D **83**, 124022 (2011)
25.12 D.M. Eardley: Black hole boundary conditions and coordinate conditions, Phys. Rev. D **57**, 2299–2304 (1998)
25.13 I. Ben-Dov: Outer trapped surfaces in Vaidya spacetimes, Phys. Rev. D **75**, 064007 (2007)
25.14 I. Bengtsson, J.M.M. Senovilla: A note on trapped surfaces in the Vaidya solution, Phys. Rev. D **79**, 024027 (2009)
25.15 I. Bengtsson, J.M.M. Senovilla: The region with trapped surfaces in spherical symmetry, its core, and their boundaries, Phys. Rev. D **83**, 044012 (2011)
25.16 R.P. Geroch, G. Horowitz: Asymptotically simple does not imply asymptotically Minkowskian, Phys. Rev. Lett. **40**, 203–206 (1978)
25.17 P. Diener: A new general purpose event horizon finder for 3-D numerical space-times, Class. Quantum Gravity **20**, 4901–4918 (2003)
25.18 R.P. Geroch, J.B. Hartle: Distorted black holes, J. Math. Phys. **23**, 680 (1982)
25.19 S. Hayward, M. Kriele: Outer trapped surfaces and their apparent horizon, J. Math. Phys. **38**, 1593 (1997)
25.20 L. Andersson, J. Metzger: The area of horizons and the trapped region, Commun. Math. Phys. **290**, 941–972 (2009)
25.21 R. Newman: Topology and stability of marginal 2-surfaces, Class. Quantum Gravity **4**, 277–290 (1987)
25.22 L. Andersson, M. Mars, J. Metzger, W. Simon: The time evolution of marginally trapped surfaces, Class. Quantum Gravity **26**, 085018 (2009)
25.23 L. Andersson, M. Mars, W. Simon: Stability of marginally outer trapped surfaces and existence of marginally outer trapped tubes, Adv. Theor. Math. Phys. **12**, 853–888 (2008)
25.24 L. Andersson, M. Mars, W. Simon: Local existence of dynamical and trapping horizons, Phys. Rev. Lett. **95**, 111102 (2005)
25.25 S. Hayward: General laws of black hole dynamics, Phys. Rev. D **49**, 6467–6474 (1994)
25.26 S.A. Hayward: Spin coefficient form of the new laws of black hole dynamics, Class. Quantum Gravity **11**, 3025–3036 (1994)
25.27 A. Ashtekar, B. Krishnan: Dynamical horizons: Energy, angular momentum, fluxes and balance laws, Phys. Rev. Lett. **89**, 261101 (2002)
25.28 A. Ashtekar, B. Krishnan: Dynamical horizons and their properties, Phys. Rev. D **68**, 104030 (2003)
25.29 E. Newman, R. Penrose: An approach to gravitational radiation by a method of spin coefficients, J. Math. Phys. **3**, 566–578 (1962)

25.30 E. Newman, T. Unti: Behavior of asymptotically flat empty space, J. Math. Phys. **3**, 891–901 (1962)
25.31 A. Komar: Covariant conservation laws in general relativity, Phys. Rev. **113**, 934 (1959)
25.32 R.P. Geroch: Multipole moments. II. Curved space, J. Math. Phys. **11**, 2580–2588 (1970)
25.33 R.O. Hansen: Multipole moments of stationary space-times, J. Math. Phys. **15**, 46–52 (1974)
25.34 R.W.S. Beig: The stationary gravitational field near spatiak infinity, Gen. Relativ. Gravit. **12**, 1003–1013 (1980)
25.35 A. Bejancu, K. Duggal: *Lightlike Submanifolds of Semi-Riemannian Manifolds and Applications* (Kluwer Academic, Dordrecht 1996)
25.36 A. Ashtekar, C. Beetle, J. Lewandowski: Mechanics of rotating isolated horizons, Phys. Rev. D **64**, 044016 (2001)
25.37 A. Ashtekar, C. Beetle, J. Lewandowski: Geometry of generic isolated horizons, Class. Quantum Gravity **19**, 1195–1225 (2002)
25.38 A. Ashtekar, S. Fairhurst, B. Krishnan: Isolated horizons: Hamiltonian evolution and the first law, Phys. Rev. D **62**, 104025 (2000)
25.39 A. Ashtekar, J. Engle, T. Pawlowski, C. van den Broeck: Multipole moments of isolated horizons, Class. Quantum Gravity **21**, 2549–2570 (2004)
25.40 E. Schnetter, B. Krishnan, F. Beyer: Introduction to dynamical horizons in numerical relativity, Phys. Rev. D **74**, 024028 (2006)
25.41 M. Jasiulek: A new method to compute quasi-local spin and other invariants on marginally trapped surfaces, Class. Quantum Gravity **26**, 245008 (2009)
25.42 J.L. Jaramillo, R.P. Macedo, P. Moesta, L. Rezzolla: Black-hole horizons as probes of black-hole dynamics I: post-merger recoil in head-on collisions, Phys. Rev. D **85**, 084030 (2012)
25.43 M. Saijo: Dynamic black holes through gravitational collapse: Analysis of multipole moment of the curvatures on the horizon, Phys. Rev. D **83**, 124031 (2011)
25.44 P. Amaro-Seoane, J.R. Gair, M. Freitag, M.C. Miller, I. Mandel, C. Cutler, S. Babak: Astrophysics, detection and science applications of intermediate- and extreme mass-ratio inspirals, Class. Quantum Gravity **24**, R113–R169 (2007)
25.45 F.D. Ryan: Accuracy of estimating the multipole moments of a massive body from the gravitational waves of a binary inspiral, Phys. Rev. D **56**, 1845–1855 (1997)
25.46 S.A. Hughes: Gravitational waves from extreme mass ratio inspirals: Challenges in mapping the space-time of massive, compact objects, Class. Quantum Gravity **18**, 4067–4074 (2001)
25.47 H. Friedrich: On the regular and the asymptotic characteristic initial value problem for Einstein's vacuum field equations, Proc. R. Soc. A **375**, 169–184 (1981)
25.48 A.D. Rendall: Reduction of the characteristic initial value problem to the Cauchy problem and its applications to the Einstein equations, Proc. R. Soc. Lond. **427**, 221–239 (1990)
25.49 H. Friedrich, A.D. Rendall: The Cauchy problem for the Einstein equations. In: *Einstein's Field Equations and Their Physical Interpretation*, Lect. Notes Phys., Vol. 540, ed. by B.G. Schmidt (Springer, Berlin, Heidelberg 2000) pp. 127–224
25.50 J. Lewandowski: Spacetimes admitting isolated horizons, Class. Quantum Gravity **17**, L53–L59 (2000)
25.51 B. Krishnan: The spacetime in the neighborhood of a general isolated black hole, Class. Quantum Gravity **29**, 205006 (2012)
25.52 I. Booth: Spacetime near isolated and dynamical trapping horizons, Phys. Rev. D **87**, 024008 (2013)
25.53 A. Ashtekar, C. Beetle, O. Dreyer, S. Fairhurst, B. Krishnan, J. Lewandowski, J. Wiśniewski: Isolated horizons and their applications, Phys. Rev. Lett. **85**, 3564–3567 (2000)
25.54 J.D. Bekenstein: Black holes and entropy, Phys. Rev. D **7**, 2333–2346 (1973)
25.55 J.M. Bardeen, B. Carter, S.W. Hawking: The four laws of black hole mechanics, Commun. Math. Phys. **31**, 161–170 (1973)
25.56 A. Ashtekar, C. Beetle, S. Fairhurst: Isolated horizons: A generalization of black hole mechanics, Class. Quantum Gravity **16**, L1–L7 (1999)
25.57 A. Ashtekar, C. Beetle, S. Fairhurst: Mechanics of Isolated Horizons, Class. Quantum Gravity **17**, 253–298 (2000)
25.58 O. Dreyer, B. Krishnan, D. Shoemaker, E. Schnetter: Introduction to isolated horizons in numerical relativity, Phys. Rev. D **67**, 024018 (2003)
25.59 S.W. Hawking, J.B. Hartle: Energy and angular momentum flow into a black hole, Commun. Math. Phys. **27**, 283–290 (1972)
25.60 E. Gourgoulhon, J.L. Jaramillo: Area evolution, bulk viscosity and entropy principles for dynamical horizons, Phys. Rev. D **74**, 087502 (2006)
25.61 T. Damour: Quelques propriétés mécaniques, électromagnétiques, thermodynamiques et quantiques des trous noirs (University of Paris, Paris 1979)
25.62 J.L. Jaramillo, M. Reiris, S. Dain: Black hole area-angular momentum inequality in non-vacuum spacetimes, Phys. Rev. D **84**, 121503 (2011)
25.63 A. Ashtekar, G.J. Galloway: Some uniqueness results for dynamical horizons, Adv. Theor. Math. Phys. **9**, 1–30 (2005)
25.64 E. Gourgoulhon, J.L. Jaramillo: A 3+1 perspective on null hypersurfaces and isolated horizons, Phys. Rep. **423**, 159–294 (2006)
25.65 I. Booth: Black hole boundaries, Can. J. Phys. **83**, 1073–1099 (2005)
25.66 A. Ashtekar, B. Krishnan: Isolated and dynamical horizons and their applications, Living Rev. Relativ. **7**, 10 (2004)

26. Gravitational Astronomy

B. Suryanarayana Sathyaprakash

This chapter is about opening the gravitational window to observe the Universe. Although the weakest of all known forces, gravity plays a dominant role in forming stars and galaxies, shaping the large-scale structure, and driving the expansion of the Universe. Gravity has so far played a passive role in our understanding. We only witness its influence indirectly by observing its effect on star light (Doppler effect, cosmological redshift, gravitational lensing, etc.). However, we are at a momentous period that could soon transform our picture of the Universe by opening the gravitational window for observational astronomy. Gravitational waves have already been critical for understanding how neutron star binaries evolve [26.1, 2]. However, we have not directly observed the waves themselves. This will change before the end of this decade when several different methods of observing gravitational waves will reach sensitivity levels at which we should finally begin to unravel some of the deepest questions in astronomy, cosmology, and fundamental physics. The chapter by van den Broeck will deal with the two latter topics. In this chapter, we will discuss what gravitational waves are (Sect. 26.2), how they interact with matter (Sect. 26.3), on-going and future projects aimed at detecting cosmic gravitational waves (Sect. 26.4), expected and speculative astronomical sources, and a list of open problems on which gravitational astronomy could shed some light (Sect. 26.5).

26.1 Background and Motivation

Astronomy began with optical telescopes. Galileo's seminal observations in ca. 1610 of Jovain moons just ≈ 30 light minutes away was a massive blow to the geocentric view and forever changed our conception of the Universe. It took another 300 years before technological advances made it possible in the early twentieth century to build telescopes that were able to detect star light outside the Milky Way and establish that we live in an expanding Universe. The nineteenth and twentieth centuries also saw another revolution in astronomy,

namely observations outside the visible part of the electromagnetic spectrum. Indeed, if our only view of the Universe was through the visible part of the electromagnetic spectrum we would not have discovered that the Universe started in a big bang as evidenced by the cosmic *microwave* background, or that some massive stars could end their lives as rapidly spinning neutron stars which emit pulses of *radio* waves, nor that black holes could power intense source of *x-rays*. These are but a small number of examples to show that stars are multimessengers that deposit quite a bit of energy in parts of the electromagnetic spectrum that would go unnoticed if all we had were optical telescopes. In fact, at different stages in their evolution stars deposit energy in infra-red, optical, UV, neutrinos, cosmic rays, radio, x-rays, gamma rays, and gravitational radiation. The only way to build a comprehensive picture of the Universe is to watch it through all of these windows and the twentieth century paved the way for making observations in all but one.

26.2 What Are Gravitational Waves?

Gravitational waves are generic to any theory of gravity that is consistent with the special theory of relativity. In any relativistic theory interactions propagate at a finite speed and gravity is no exception. In fact, many attempts were made in the past to incorporate the finite speed of gravity (most notably by *Laplace* [26.3]), and the existence of gravitational waves was envisaged by Poincaré well before Einstein developed his general theory of relativity. In this section we will provide a heuristic picture, at the risk of a lack of rigor, of what gravitational waves are, highlighting only the key results and leaving out all intermediate steps of calculations.

26.2.1 The Newtonian Picture and Maxwell's Equations

In Newtonian gravity the gravitational potential φ generated by mass density ρ is given by the Poisson equation

$$\nabla^2\varphi = 4\pi G\rho\,, \tag{26.1}$$

where G is Newton's gravitational constant. We can think of this equation as the *limit* of the wave equation in which the speed v of the interaction is infinite

$$-\frac{1}{v^2}\frac{\partial^2\varphi}{\partial t^2} + \nabla^2\varphi = 4\pi G\rho\,. \tag{26.2}$$

In the Newtonian picture, therefore, gravity propagates at infinite speed and the potential exhibits no wave-like properties. If this were true, it would have been possible to build a *gravity telegraph* that would, in principle, transmit signals at infinite speed, which would be inconsistent with special relativity.

On the contrary, Maxwell's equations of electrodynamics are, in the Lorenz gauge [26.4] (most textbooks wrongly attribute these gauge conditions to H.A. Lorentz; in reality it was L. Lorenz who found them first, see, for instance, [26.5]), explicit wave equations in vector and scalar potentials, $\boldsymbol{A}$ and ϕ, respectively

$$\begin{aligned} -\mu\epsilon\frac{\partial^2\boldsymbol{A}}{\partial t^2} + \nabla^2\boldsymbol{A} &= -\mu\boldsymbol{J}\,,\\ -\mu\epsilon\frac{\partial^2\phi}{\partial t^2} + \nabla^2\phi &= -\frac{\rho}{\epsilon}\,. \end{aligned} \tag{26.3}$$

Here $\boldsymbol{J}$ and ρ are the current and charge densities, ϵ is the permittivity of the medium and μ its permeability. The general solution to these equations consists of not only the Coulomb and induction fields that fall off with distance as r^{-2} but also radiative fields that fall off as r^{-1} and propagate at a *finite* speed $1/\sqrt{\epsilon\mu}$. Maxwell noticed that in vacuum (permittivity $\epsilon_0 = 8.854\times10^{-12}\,\mathrm{F\,m^{-1}}$ and permeability $\mu_0 = 4\pi\times10^{-7}\,\mathrm{H\,m^{-1}}$) the speed $c = 1/\sqrt{\mu_0\epsilon_0} = 3.00\times10^8\,\mathrm{m\,s^{-1}}$ is the same as the speed of light, which prompted him to propose that light is a form of electromagnetic waves – a truly remarkable feat.

26.2.2 Einstein's Gravity and Gravitational Waves

Einstein is believed to have been greatly influenced by the consistency of Maxwell's equations with special relativity in developing his new theory of gravity, in which gravitational interaction has to necessarily propagate at a finite speed. General relativity differs from Newtonian theory in other respects too. Firstly, all forms of matter and energy, including stresses and pressures that could initially work against gravity, are sources of the field.

This has the consequence that when objects reach a certain *compactness* the very stresses that are needed for their stability will facilitate gravitational collapse, leading to the formation of a black hole. (Compactness κ is a measure of an object's gravitational radius $R_G = GM/c^2$ relative to its actual radius, i.e., $\kappa \equiv R_G/R = GM/c^2R$, where M is the object's mass.)

Secondly, instead of a single scalar field φ, general relativity contains ten potentials, the components of a symmetric second rank tensor $g_{\mu\nu}$, corresponding to the background metric of spacetime. Far away from an isolated source the geometry is *close* to Minkowski spacetime. It is then possible to write the metric such that it is only a small perturbation of flat spacetime

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}\,, \quad \|h_{\mu\nu}\| \ll 1\,. \tag{26.4}$$

When only linear terms are kept in the *metric perturbation* $h_{\mu\nu}$ Einstein equations simplify to wave equations

$$-\frac{1}{c^2}\frac{\partial^2 \bar{h}_{\mu\nu}}{\partial t^2} + \nabla^2 \bar{h}_{\mu\nu} = 16\pi G T_{\mu\nu}\,, \tag{26.5}$$

where $\bar{h}_{\mu\nu} = h_{\mu\nu} - (h/2)\eta_{\mu\nu}$ is called the *trace-reverse* of $h_{\mu\nu}$ and $T_{\mu\nu}$ is the energy–momentum tensor of the source of the gravitational field. As in the case of Maxwell's equations, here too the general solution consists of a Newtonian field that falls off as r^{-2} and a radiative solution that drops off as r^{-1}. Thus, Einstein's theory admits wave-like solutions that travel outward from their source at the speed of light. (Owing to their very weak interaction, gravitational waves are not easily dispersed or attenuated by the medium in which they travel. As a result there is no need to introduce the concepts of permittivity and permeability of a medium for gravitational waves.)

Thirdly, general relativity is a generally covariant theory and so Einstein equations are the same in all coordinate frames. Together with the fact that four of the ten Einstein equations are constraint equations, general covariance implies that there are only *two* dynamical degrees of freedom in the theory. Similarly, weak gravitational fields and gravitational waves are also described by two degrees of freedom, or two polarizations. To specify these two degrees of freedom, it is necessary to choose a specific gauge and coordinate system. A convenient gauge choice is the one that keeps only the transverse part of the metric perturbation nonzero (i.e., $\bar{h}_{\mu\nu}k^\nu = 0$, where k^ν is the wave vector) and makes it traceless (i.e., $\bar{h}^{\mu}{}_{\mu} = 0$). For a wave traveling along z axis these conditions imply

$$g_{\mu\nu} = \begin{pmatrix} -c^2 & 0 & 0 & 0 \\ 0 & 1 + h_+\left(t - \frac{z}{c}\right) & h_\times\left(t - \frac{z}{c}\right) & 0 \\ 0 & h_\times\left(t - \frac{z}{c}\right) & 1 - h_+\left(t - \frac{z}{c}\right) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{26.6}$$

where $h_+ \equiv \bar{h}^{\mathrm{TT}}_{xx} = -\bar{h}^{\mathrm{TT}}_{yy}$ is called the *plus* polarization, $h_\times \equiv \bar{h}^{\mathrm{TT}}_{xy}$ is called the *cross* polarization, and the symbol TT is used to denote that the metric components are computed in the *transverse-traceless* gauge.

Gravitational waves are sometimes referred to as *ripples in the very fabric of spacetime*. This phraseology stems from the fact that the components of the curvature tensor themselves exhibit oscillatory behavior in the presence of gravitational waves. Moreover, the effect of gravitational waves is in many ways similar to that of tidal gravitational forces but with one important difference: the tidal field due to radiation falls off as r^{-1}, while tidal fields of stationary fields fall off as r^{-3}.

In Schwarzschild geometry, in a freely falling frame, the nonzero independent components (a caret on components indicates that this is a *freely falling* system of coordinates and not the usual Schwarzschild coordinates) of the Riemann tensor are proportional to M/r^3 [26.6]

$$R^{\hat{r}}{}_{\hat{t}\hat{r}\hat{t}} = -R^{\hat{\theta}}{}_{\hat{\varphi}\hat{\theta}\hat{\varphi}} = -\frac{2M}{r^3}\,,$$
$$R^{\hat{\theta}}{}_{\hat{t}\hat{\theta}\hat{t}} = R^{\hat{\varphi}}{}_{\hat{t}\hat{\varphi}\hat{t}} = -R^{\hat{\theta}}{}_{\hat{r}\hat{\theta}\hat{r}} = -R^{\hat{\varphi}}{}_{\hat{r}\hat{\varphi}\hat{r}} = \frac{M}{r^3}\,,$$

where M is the mass of the black hole and r is the radial distance. This is reminiscent of the tidal field in Newtonian gravity. The effect of this field is to cause tidal deformation of nearby (radial) geodesics, stretching the distance between them in the radial direction and squeezing in the transverse direction due to the opposite signs of $R^{\hat{r}}{}_{\hat{t}\hat{r}\hat{t}}$ and $R^{\hat{\theta}}{}_{\hat{t}\hat{\theta}\hat{t}}$. It is important to note that this field drops off with distance as r^{-3}, and so distant objects have negligible tidal fields. However, this is not so in the case of gravitational radiation.

For a plane wave traveling in the z direction, the nonzero independent components of the curvature ten-

sor in the TT gauge are

$$R^x{}_{0x0} = -\frac{1}{2}\ddot{h}^{\mathrm{TT}}_{xx} ,$$
$$R^y{}_{0x0} = -\frac{1}{2}\ddot{h}^{\mathrm{TT}}_{xy} ,$$
$$R^y{}_{0y0} = -R^x{}_{0x0} , \tag{26.7}$$

from which we can see that the components of the curvature tensor fall off as inverse of the distance from the source as $h \propto r^{-1}$. Even though the stationary part of the field falls off very rapidly, the radiative part of the field decreases only as inverse of the distance. This explains why gravitational waves from a system can be detected although the stationary part of the field might be completely negligible. Moreover, since $R^y{}_{0y0} = -R^x{}_{0x0}$ the effect of gravitational waves is also to cause a tidal deformation of geodesics. This property of the waves often dictates the design of a gravitational wave detector.

26.2.3 Gravitational Wave Luminosity

As stated in Sect. 26.1, for our description of the Universe gravity plays a key role (formation of stars, expansion of the Universe, etc.). Under most circumstances gravitating systems are nearly stationary and so emit a negligible amount of gravitational radiation. However, one expects relativistic sources to be very powerful emitters of gravitational waves.

In electrodynamics, one of the most common and powerful ways of relating the nature of radiation from a system to the geometry and dynamics of the source is the method of multipole expansion. The expansion results in a series that consists of monopole, followed by dipole, quadrupole, octupole, etc. A time-varying monopole, for example a charged sphere whose radius oscillates in and out, emits no radiation, a result that follows from the conservation of charge [26.7]. All other time-varying multipoles can, in general, give rise to electromagnetic radiation. The luminosity of the radiation due to the first three multipoles is [26.8, 9]

$$\mathcal{L} = \frac{1}{4\pi\epsilon_0 c^3}\left[\frac{2}{3}\ddot{d}_k\ddot{d}^k + \frac{2}{3}\ddot{\mu}_k\ddot{\mu}^k + \frac{1}{20c^2}\dddot{\mathcal{D}}_{kl}\dddot{\mathcal{D}}^{kl}\right], \tag{26.8}$$

where $\boldsymbol{d} = \sum e\boldsymbol{r}$ and $\boldsymbol{\mu} = \sum (e/2c)(\boldsymbol{r}\times\boldsymbol{v})$ are the electric and magnetic dipole moments, respectively, and $\mathcal{D}_{kl} = \sum e(r_k r_l - \frac{1}{3}\delta_{kl}r^2)$ is the electric quadrupole moment. (In these expressions, the sum is over all charges in the system, e, $\boldsymbol{r}$, and $\boldsymbol{v}$ are the charges, positions, and velocities of the particles.) The multipole expansion assumes that the size of the system is much smaller than the wavelength of radiation, in which case higher-order terms can be neglected. As we can see, the electric dipole is the most dominant term in the above expansion and it is the most efficient way to produce electromagnetic radiation. The next multipoles of importance are the magnetic dipole and electric quadrupole, each being smaller than the dipole by a factor v^2/c^2, where v is the typical speed of particles. (The similarity of the first and second terms is somewhat misleading; note that the magnetic moment has a factor of v/c in its definition, and so the contribution of the second term is $\mathcal{O}(v^2/c^2)$ smaller than the first term.) Thus, in the slow-motion approximation (another assumption made in deriving (26.8) for luminosity) the magnetic dipole and electric quadrupole contributions are very small.

A multipole expansion of the radiative solutions can be carried out also for gravitational radiation generated by a system of masses in motion. (Due to the nonlinear nature of general relativity, the full treatment of the problem of wave generation requires the application of a number of different approximation techniques beyond the multipole expansion. These include the post-Minkowskian approximation, an expansion in G; post-Newtonian approximation, an expansion in v/c; long wavelength approximation, an expansion in size a of the source; and far field expansion, an expansion in $1/r$, where r is the field point. See [26.10] for a recent review.) Just as in electromagnetism, here too the monopole term is absent – a result that follows from the conservation of mass. However, in gravity there is no dipole radiation either as the time derivative of the mass dipole is nothing but linear momentum and so its time-derivative will be zero for an isolated system of masses. The next most important terms are the analogs of the magnetic dipole (called *current* dipole) and electric quadrupole (called *mass quadrupole*). The current dipole is nothing but the angular momentum $\boldsymbol{L} = \sum m\boldsymbol{r}\times\boldsymbol{v}$. Conservation of angular momentum implies that there will be no current dipole radiation either [26.9].

The most dominant multipole radiation in gravity is a time-varying mass quadrupole. The luminosity in gravitational waves from a system with mass density $\rho(\boldsymbol{r}, t)$ is given by

$$\mathcal{L} = \frac{G}{5c^5}\dddot{Q}_{ij}\dddot{Q}^{ij} ,$$
$$Q_{ij} = \int \rho\left(r_i r_j - \frac{1}{3}\delta_{ij}r^2\right)\mathrm{d}^3x . \tag{26.9}$$

This is the famous quadrupole formula and one of the earliest formulas to be derived in the context of gravitational wave generation [26.11] (see also [26.8]). The quantity Q_{ij} is called the *reduced* quadrupole moment to distinguish it from other similar quantities that appear in other areas of physics. Several interesting points emerge from the quadrupole formula:

1. For a spherically symmetric mass distribution the reduced quadrupole moment is identically zero. Therefore, it is only nonspherical accelerations of masses that can produce any radiation. In general, axisymmetric motion can produce radiation; however, an axisymmetric body rotating about its symmetry axis does not.
2. As in the case of electric quadrupole, the luminosity is suppressed by a factor of c^5. Consequently, only systems with relativistic velocities can be expected to have appreciable emission of gravitational radiation.
3. Since the luminosity depends on the third derivative of the quadrupole, it is not enough for a system to have a large quadrupole in order to be a good source of gravitational radiation; the quadrupole must change rapidly to efficiently convert mechanical energy into gravitational radiation. For example, a system could have a very large quadrupole moment if it has a large size, but gravitationalwaves will be weak unless the system is rapidly accelerating.
4. Luminosity grows rapidly as a function of frequency ω or equivalently compactness κ and velocity v in the system. Purely from dimensional analysis (and assuming that only two terms in the sum in (26.9) survive) we see that for a system whose frequency scale is ω and physical scale is R, luminosity grows as $\mathcal{L} \approx 2GM^2R^4\omega^6/(5c^5)$. Furthermore, for self-gravitating systems $\omega^2 \approx GM/R^3$ (equivalently the velocity v in the system is $v^2 \approx GM/R$) and so

$$\mathcal{L} \approx \frac{2c^5}{5G}\left(\frac{GM\omega}{c^3}\right)^{10/3} \approx \frac{2c^5}{5G}\kappa^5 \approx \frac{2c^5}{5G}\left(\frac{v}{c}\right)^{10} .$$

This shows that systems have to be very compact (i. e., large κ) in order for the radiation to be appreciable. Equivalently, luminosity grows as a very steep power of the nonspherical velocity in the system. The constant $c^5/G \simeq 3.6 \times 10^{52}\,\mathrm{J\,s^{-1}} \approx 10^{26} L_\odot$, is the factor needed to go from the dimensionless luminosity κ^5, to one that has physical dimensions. This is an enormous luminosity, at least a factor 10^3 larger than the luminosity in visible light of all the stars in the Universe. Since it is multiplied by the fifth power of compactness, most gravitational wave sources never reach this luminosity. However, in the case of binary black holes, for which the compactness can be of order unity, the luminosity does come close to this stupendous value independent of the total mass of the system, although the duration over which the system has this large luminosity is greater for more massive systems.

All of the above go on to show why gravitational waves are not easy to produce: the source must be relativistic and compact with a strongly time-varying quadrupole moment. Evidently, all of these conditions are related to each other in a self-gravitating dynamical system (e.g., nonspherical collapse of a star or a tight binary consisting of compact stars), and so when any one of the conditions is met there is a good chance others are met too. So we should expect prominent sources of gravitational waves to be compact, relativistic sources.

How long can a source of gravitational radiation last? It depends on the total amount of energy that is available to convert into radiation. For example, in a binary system of stars on a circular orbit of size R the available energy is roughly $-G\nu M^2/2R$, where $M = m_1 + m_2$ is the total mass of the system and $\nu = m_1 m_2/M^2$ is the symmetric mass ratio. A source of luminosity $\mathcal{L}$, with energy $\mathrm{d}E$ will last for a time $\mathrm{d}t = \mathrm{d}E/\mathcal{L}$. The time t_C it takes for the system's energy to change from E_i to E_f (size R_i to $R_\mathrm{f} \ll R_\mathrm{i}$) is

$$t_\mathrm{C} = \int_{E_\mathrm{i}}^{E_\mathrm{f}} \frac{\mathrm{d}E}{\mathcal{L}} = \int_{R_\mathrm{i}}^{R_\mathrm{f}} \frac{1}{\mathcal{L}}\frac{\mathrm{d}E}{\mathrm{d}R}\,\mathrm{d}R \simeq \frac{5G}{256c^3}\frac{M}{\nu}\kappa_\mathrm{i}^{-4} , \tag{26.10}$$

where $\kappa_\mathrm{i} = GM/(c^2R_\mathrm{i})$ is the initial compactness of the source. This equation shows that for noncompact sources ($\kappa_\mathrm{i} \ll 1$), the gravitational radiation damping operates on extremely large time scales. Stellar mass binaries spend millions of years with large orbital periods of $\approx$ hours or days, but will only last for a few days or hours when they reach orbital periods of $\approx$ seconds.

Let us finally note that the quadrupole formula is the lowest-order post-Newtonian approximation which assumes that:

a) The gravitational field inside the source is weak
b) The motion inside the source is slow
c) The size of the system is small compared to the wavelength of the emitted waves.

The quadrupole formula has been vindicated by radio observations of the Hulse–Taylor binary [26.12], a system comprising of two neutron stars in a tight orbit. Given the gravitational wave luminosity of the binary and its energy E, one can use the energy balance equation $\mathcal{L} = -\frac{dE}{dt}$ to compute the orbital evolution of the system. The energy balance equation simply states that the gravitational wave luminosity results in loss in binding energy of the system. The quadrupole formula predicts that the binary should emit gravitational waves and this loss of energy should cause the binary orbit to shrink. The observed rate of change of the period $\dot{P}_b$ agrees with the prediction of the quadrupole formula, $\dot{P}_{\rm GR} = -2.402531 \pm 0.000014 \times 10^{-12}$, to better than 0.2%: $\dot{P}_b/\dot{P}_{\rm GR} = 0.997 \pm 0.002$ [26.1, 2].

26.2.4 Wave Amplitudes in Terms of Source Moments

What is the relationship between the luminosity in gravitational waves to the wave amplitudes? How are the two gravitational wave polarizations related to the multipole moments of the source? These are important questions because they would help in calculating the magnitude of the amplitudes from source luminosities but also, more importantly, in inferring the dynamics of a source by observing the radiation that it emits.

The wave amplitudes $h_{ij}^{\rm TT}$ are related to the quadrupole tensor Q_{ij} of the source by [26.10]

$$h_{ij}^{\rm TT} = \frac{2G}{c^4 r}\mathcal{P}_{ijab}\frac{d^2 Q_{ab}}{dt^2} ,$$
$$Q_{ij} = \int \rho \left(x_i x_j - \frac{1}{3}(x^k x_k)\delta_{ij} \right) d^3 x , \qquad (26.11)$$

where $r = |\boldsymbol{r}|$ is the distance to the source, ρ is the source density, and $\mathcal{P}_{ijab} = \mathcal{P}_{ia}\mathcal{P}_{jb} - \frac{1}{2}\delta_{ij}\mathcal{P}_{ij}\mathcal{P}_{ab}$ is the TT projection operator, where $\mathcal{P}_{ij} = \delta_{ij} - n_i n_j$ is the operator that projects orthogonal to the vector $\hat{n} \equiv \boldsymbol{r}/r$.

The combination $c^4/G \simeq 1.2 \times 10^{44}$ N has dimensions of force. One can think of the inverse of this quantity (recall that the coupling of the stress–energy tensor of matter to the Einstein tensor in the field equations is $8\pi G/c^4$) as a measure of the coupling of accelerated motion of bodies to the curvature of spacetime. A force of this order of magnitude is required to cause changes of order unity in the geodesics. This illustrates that spacetime is extremely stiff and enormous forces are required to change its curvature.

Furthermore, the second derivative of the quadrupole moment $\ddot{Q}_{ij}$ has dimensions of energy; it is the energy in the nonspherical motion of the system. For spherically symmetric motions the quadrupole moment Q_{ij} is identically zero. For a system in which the entire mass M of the system is in nonspherical motion (a situation that occurs in BBH), components of $\ddot{Q}_{ij}$ are of order $\omega^2 M R^2$, where, as before, ω is the frequency scale of the source. Additionally, for self-gravitating sources, the frequency is related to its size R via Kepler's law: $\omega^2 \approx GM/R^3$, giving the maximum one can expect for the amplitude to be

$$h \lesssim \frac{GM}{c^2 R} \times \frac{GM}{c^2 r} = \kappa\phi_{\rm ext} , \quad \text{where} \quad \phi_{\rm ext} = \frac{GM}{c^2 r} . \qquad (26.12)$$

Thus, for a source at a given distance, the amplitude of the waves is the greatest for the most compact source, when it is equal to the external gravitational potential $\phi_{\rm ext}$ of the source at the observation point. Black holes and neutron stars are the most compact objects in the Universe and interactions involving them are the primary sources of gravitational waves.

For a binary of total mass $M = 10 M_\odot$, at a distance of 100 Mpc, the amplitude of the source when the binary is about to coalesce (i. e., $\kappa \approx \frac{1}{2}$) is $h \approx 10^{-21}$. Recall that at this stage the luminosity in gravitational waves would be $\mathcal{L} \approx 4 \times 10^{50}\,\mathrm{J\,s^{-1}}$, a very bright source indeed. Yet the amplitude is not so large due to the weak coupling of gravitational waves to matter. It is useful to obtain a rough idea of how the amplitude of gravitational waves is related to the source's luminosity. Comparing (26.9) and (26.11) and writing $\dot{h} = \omega h = 2\pi f h$, where f is the frequency of gravitational waves, we observe that, as an order of magnitude

$$\mathcal{L} \approx \frac{c^3}{20G}(r\dot{h})^2 ,$$
$$h^2 \approx \frac{20G}{c^3}\frac{\mathcal{L}}{\omega^2 r^2} \Rightarrow h \approx \sqrt{\frac{5G}{c^3}\frac{\Delta E}{\Delta t}}\frac{1}{\pi f r} , \qquad (26.13)$$

where ΔE is the energy in gravitational waves emitted over a time Δt.

26.3 Interaction of Gravitational Waves with Light and Matter

Understanding the basic effect of the waves on time-like and null geodesics is essential in appreciating how different schemes designed to detect gravitational waves work, but also necessary in discovering new methods of detection. In this section we will discuss how gravitational waves interact with light and matter.

Universality of gravitation and the *equivalence principle* imply that it is impossible to distinguish between noninertial reference frames from gravitational fields in a small local neighborhood in spacetime: no experiment restricted to infinitesimally small time and length scales can differentiate between accelerated reference frames from gravitational fields (see, e.g., [26.6]). For instance, no local experiment would detect the presence of the Earth's gravity in a freely falling lift. One way to infer the presence of such a field is to compare the frequency of a standard source of light as it propagates from one point to another. As light *climbs* up a gravitational potential it is redshifted. Another way to detect gravitational fields is to watch two nearby freely falling particles. After some time, the two particles will be seen to approach each other. The effect of gravitational waves on light beams and free test masses is no different from the effect of gravity itself. We will discuss the two physical effects in the context of gravitational waves. To this end it will be useful to recall the form of the metric for weak gravitational waves and associated symmetries. The metric of a plane gravitational wave traveling in the z direction is given by (26.6)

$$\begin{aligned} ds^2 = &-c^2 dt^2 + \left[1 + h_+\left(t - \frac{z}{c}\right)\right] dx^2 \\ &+ \left[1 - h_+\left(t - \frac{z}{c}\right)\right] dy^2 \\ &+ 2h_\times\left(t - \frac{z}{c}\right) dx dy + dz^2 . \end{aligned} \quad (26.14)$$

In *null coordinates*, defined by $\xi = ct + z$ and $\chi = ct - z$, the metric takes the form

$$\begin{aligned} ds^2 = &-d\xi\, d\chi + [1 + h_+(\chi)]\, dx^2 \\ &+ [1 - h_+(\chi)]\, dy^2 \\ &+ 2h_\times(\chi)\, dx dy . \end{aligned}$$

This metric is independent of the coordinates (ξ, x, y) and so there are three Killing vectors $K_1 = \partial/\partial\xi$, $K_2 = \partial/\partial x$, and $K_3 \partial/\partial y$, with components in the (t, x, y, z) coordinates given by

$$\begin{aligned} K_1^\mu &= (c, 0, 0, 1) , \\ K_2^\mu &= (0, 1, 0, 0) , \\ K_3^\mu &= (0, 0, 1, 0) . \end{aligned} \quad (26.15)$$

We will use these results in the remainder of this section to discuss the effect of gravitational waves on null and time-like geodesics.

26.3.1 Doppler Modulation of Light in the Presence of Gravitational Waves

The path of light in a gravitational field is described by null geodesics. This is true even when the background spacetime is that of a gravitational wave. Just as in stationary gravitational fields, we can use the Doppler effect of the field on null geodesics to detect gravitational waves.

Let us consider a beam of light in the field of a gravitational wave propagating in the z direction. The beam is sent from the emitter located at the origin of the coordinate system to a receiver a distance L away from the emitter, as in Fig. 26.1. The directions of the light beam and gravitational wave define a plane which we take

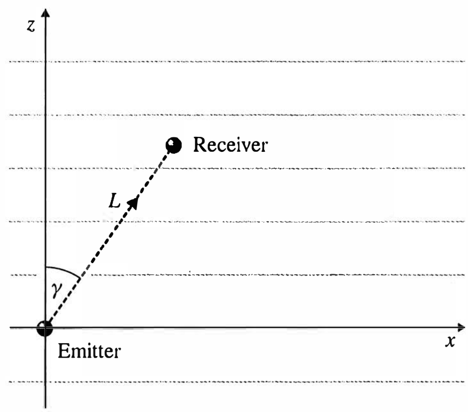


Fig. 26.1 A light beam (*dashed line*) making an angle γ with the z axis and traveling in the xz plane is sent from the emitter to a receiver located at a distance L from the emitter. The frequency of light at the receiver is Doppler modulated relative to the emitter as it travels in the field of a gravitational wave propagating in the z direction (whose wave fronts are shown parallel to the x axis)

to be the xz plane. Furthermore, let us assume that the wave consists of only the plus polarization, i. e., $h_+ \neq 0$ and $h_\times = 0$. In this case, the metric simplifies to

$$\mathrm{d}s^2 = -c^2\,\mathrm{d}t^2 + \left[1 + h_+\left(t - \frac{z}{c}\right)\right]\mathrm{d}x^2 + \left[1 - h_+\left(t - \frac{z}{c}\right)\right]\mathrm{d}y^2 + \mathrm{d}z^2\,. \quad (26.16)$$

Although we have assumed a specific polarization and direction for the propagation of the waves, the final result can be written in a covariant form.

Let the light beam make an angle γ with the z axis. In flat spacetime such a light beam will be described by a null vector $U^\mu = \nu(1, c\sin\gamma, 0, c\cos\gamma)$, where ν is the frequency of light. U^μ parallel transported along itself defines the path of light and, in flat spacetime, both ν and γ remain fixed as the beam propagates. In a curved spacetime, however, neither will remain fixed. Following *Estabrook* and *Wahlquist* et al. [26.13], we will compute how the frequency of light is modulated.

Let us denote by ν_{E} and ν_{R} frequencies, by γ_{E} and γ_{R} angles, and by V^μ_{E} and V^μ_{R} null vectors, at the emitter and receiver, respectively. To linear order in the metric perturbation $h_{\alpha\beta}$, a null vector V^μ in the perturbed spacetime is related to the flat spacetime null vector U^μ by $V^\mu = U^\mu - \frac{1}{2}\eta^{\mu\alpha}h_{\alpha\beta}U^\beta$. Using this relation it is easy to see that

$$V^\mu_{\mathrm{E}} = \nu_{\mathrm{E}}\left(1, c\sin\gamma_{\mathrm{E}}\left(1 - \frac{1}{2}h_+(t_{\mathrm{E}})\right), 0, c\cos\gamma_{\mathrm{E}}\right)$$

$$V^\mu_{\mathrm{R}} = \nu_{\mathrm{R}}\left(1, c\sin\gamma_{\mathrm{R}} \times \left(1 - \frac{1}{2}h_+\left(t_{\mathrm{R}} - \frac{L\cos\gamma}{c}\right)\right), 0, c\cos\gamma_{\mathrm{R}}\right),$$

where t_{E} and $t_{\mathrm{R}} = t_{\mathrm{E}} + L/c$ are the times when the beam leaves the emitter and is received at the receiver, respectively. Note, however, that the gravitational field at the receiver is evaluated *not* at time t_{R} but at an earlier time $t_{\mathrm{R}} - L\cos\gamma/c$; the gravitational wave phase does not quite catch up with the beam as they travel in different directions (for γ one can use either γ_{E} or γ_{R}, as they differ only by order h).

The null vectors at the emitter and receiver are related by parallel transport and owing to the existence of Killing vectors (cf. (26.15)), we have three conserved quantities: $\phi_i \equiv g_{\mu\nu}V^\mu K^\nu_i$, for $i = 1, 2, 3$. This means that ϕ_i computed at the emitter is the same as that computed at the receiver. Due to our choice of geometry ϕ_3 is identically zero. For $i = 1, 2$ we have

$$\phi_{1\mathrm{E}} = \phi_{1\mathrm{R}} \Rightarrow \nu_{\mathrm{E}}(1 - \cos\gamma_{\mathrm{E}}) = \nu_{\mathrm{R}}(1 - \cos\gamma_{\mathrm{R}})$$

$$\phi_{2\mathrm{E}} = \phi_{2\mathrm{R}} \Rightarrow \nu_{\mathrm{E}}\left(1 + \frac{1}{2}h_+(t_{\mathrm{E}})\right)\sin\gamma_{\mathrm{E}} = \nu_{\mathrm{R}}\left(1 + \frac{1}{2}h_+\left(t_{\mathrm{R}} - \frac{L\cos\gamma}{c}\right)\right)\sin\gamma_{\mathrm{R}}\,.$$

One can eliminate γ_{R} from the above equations and solve for the Doppler shift in the beam caused by the wave. Again keeping only terms to linear order in h we find

$$\frac{\nu_{\mathrm{R}} - \nu_{\mathrm{E}}}{\nu_{\mathrm{E}}} = \frac{1 + \cos\gamma}{2} \times \left[h_+(t) - h_+\left(t - \frac{L\cos\gamma}{c}\right)\right], \quad (26.17)$$

where for γ one can use either γ_{E} or γ_{R}. Except when the beam travels in the same direction as the wave (i. e., $\gamma = 0, \pi$), Doppler modulation of light can be used to detect gravitational waves. This is the principle of operation of laser interferometers (on the ground and in space) and PTA.

An alternative interpretation of (26.17) is that the rate at which the clocks tick at the receiver is different from that at the emitter. By writing the rate at which the clocks at the emitter and receiver tick as $\Delta t_{\mathrm{E}} = \nu_{\mathrm{E}}^{-1}$ and $\Delta t_{\mathrm{R}} = \nu_{\mathrm{R}}^{-1}$ we have (to first order in h)

$$\frac{\Delta t_{\mathrm{R}}}{\Delta t_{\mathrm{E}}} = 1 + \frac{1 + \cos\gamma}{2} \times \left[h_+\left(t - \frac{L\cos\gamma}{c}\right) - h_+(t)\right]. \quad (26.18)$$

In the presence of gravitational waves clocks at the emitter and receiver tick at different rates. One can detect a passing gravitational wave by comparing the rate at which light pulses, sent at a constant rate as measured by a clock at the emitter, arrive as measured by a local clock at the receiver. Admittedly, the variation in the arrival rate is very small. It depends on the phase of the wave when the beam arrives at the receiver compared to its phase when the beam leaves the emitter, which can at best be the amplitude of the wave. The best clocks today are stable to a few parts in 10^{16} [26.14], which sets the sensitivity of a detector that uses two clocks to $h \approx 10^{-15}$. This is the sensitivity of a PTA (see below).

Instead of comparing the clocks at the emitter and receiver one can simply look at the arrival times of pulses that are sent by the emitter and then reflected

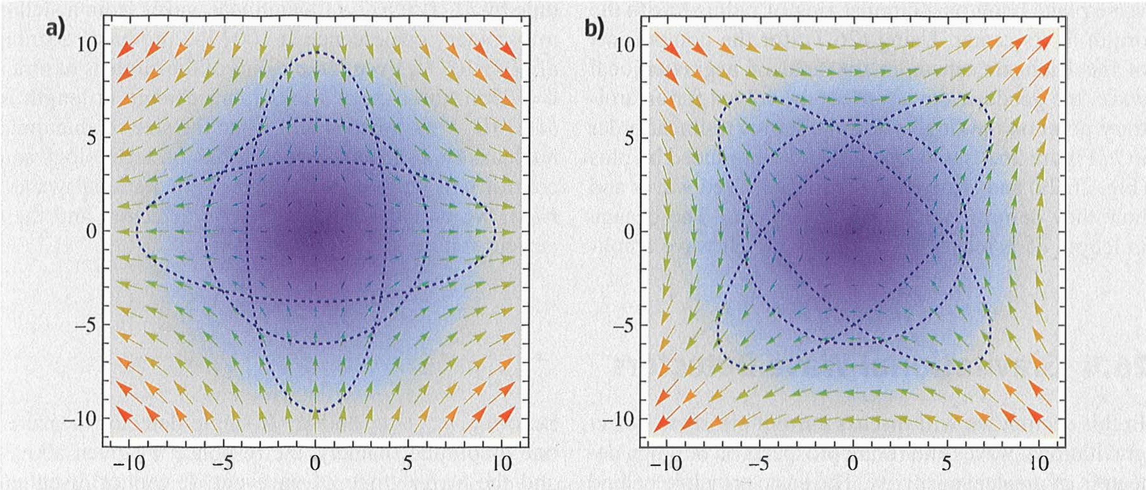


Fig.26.2a,b The diagram shows the force field and the response of a ring of free particles due to waves of plus ((**a**), $h_+ = 1.6\cos\omega t$, $h_\times = 0$) and cross ((**b**), $h_+ = 0$, $h_\times = 1.6\cos\omega t$) polarizations passing perpendicular to the plane of the paper. The force field is shown when the phase of the wave is $\omega t = 2n\pi$, n being an integer. The strain $\delta\ell/\ell$, where $\delta\ell$ is the change in the size ℓ of the ring, is equal to the amplitude h of the wave: $\delta\ell/\ell \approx h/2$. Over one gravitational wave cycle, the ring deforms from one ellipse through the circle to the other ellipse and back to the first ellipse. The deformation preserves the area

by the receiver back to the emitter. It is easy to work out that the rate at which the light pulses return to the emitter is

$$\frac{\Delta t_{\text{return}}}{\Delta t_{\text{E}}} = 1 + \frac{1}{2}\left[(1-\cos\gamma)h_+\left(t+\frac{2L}{c}\right) + 2\cos\gamma h_+\left(t+\frac{L(1-\cos\gamma)}{c}\right) - (1+\cos\gamma)h_+(t)\right]. \tag{26.19}$$

Sensitivity of even this arrangement would be limited by how stable the clocks are. Interferometry avoids this problem by comparing the round trip travel time of light beams in two orthogonal directions; the round trip travel time in one arm of the interferometer is used as a reference clock against which the round trip travel time along the other arm is compared. This is the basic principal of operation of interferometric gravitational wave detectors.

26.3.2 Geodesic Deviation Equation

The foregoing discussion was focussed on the effect of gravitational waves on null geodesics. We will now consider the effect of gravitational waves on free test masses. To this end it is useful to ask how an oscillating Riemann curvature tensor would affect the motion of free particles. This is answered by the geodesic deviation equation. The vector ξ^μ connecting a reference geodesic and its neighbor obeys the geodesic equation

$$\frac{\mathrm{d}^2\xi^\mu}{\mathrm{d}\tau^2} = -R^\mu{}_{\alpha\nu\beta}U^\alpha\xi^\nu U^\beta\,, \tag{26.20}$$

where τ is the proper time and U^α is the four-velocity along the reference geodesic. In flat space-time, $R^\mu{}_{\alpha\nu\beta} = 0$ and so $\mathrm{d}\xi^\mu/\mathrm{d}\tau = \text{const.}$ If initially $\mathrm{d}\xi^\mu/\mathrm{d}\tau = 0$, then it will continue to remain zero since there is no acceleration. Thus, the proper distance between two neighboring geodesics, that are initially parallel, will remain constant. In the field of a gravitational wave the curvature tensor oscillates about a mean value of zero. When the curvature is positive, geodesics become focussed (because of the negative sign on the right-hand side of (26.20)) and so the proper distance decreases; when the curvature becomes negative, geodesics diverge and the proper distance increases from its mean value. Thus, the effect of the waves is to cause the proper distance between two neighboring geodesics to oscillate about its mean value.

To illustrate the effect more clearly, let us suppose a gravitational wave is traveling in the z direction and let us consider a system of free particles in the trans-

verse plane lying on a circular ring of radius ℓ with the origin as its center. Using (26.7), for the components of the Riemann tensor in the field of a gravitational wave, the geodesic deviation can be solved for an arbitrary point on the ring by keeping terms to linear order in h. Figure 26.2 shows the tidal field produced by plus (Fig. 26.2a) and cross (Fig. 26.2b) polarized waves and how they deform a ring of free particles. The change in length $\delta\ell$ is related to the gravitational wave amplitude by $\delta\ell/\ell \approx h/2$. Gravitational waves from a stellar mass binary coalescence at 100 Mpc produces a strain $\delta\ell/\ell \approx 10^{-21}$. Even if the radius of the circle is of order $\ell = 2$ km, the expected maximum change in length is $\delta\ell \approx 10^{-18}$ m, much smaller than the size of an atomic nucleus. This shows that high precision technology and control will be required to detect gravitational waves. Next, we will discuss the various detectors and their current status.

26.4 Gravitational Wave Detectors

In this section we will discuss current efforts to detect gravitational waves and future prospects on building detectors of greater sensitivity. The basic principle behind any detector is to either measure the strain in space caused by passing gravitational waves or to monitor the time of flight of light beams as they traverse the variable curved spacetime geometry produced by gravitational waves. These principles, of course, are really based on one and the same effect but the operation of different detectors is more readily understood by invoking one rather than the other.

Before we discuss the various efforts to build detectors, let us first note that most gravitational wave detectors have quite a good sensitivity to sources over most of the sky. Naturally, they are not good in estimating the sky position of a source. In fact, a single detector yields only a certain combination of the two polarizations called the antenna response $h^A(t)$, where A is an index representing detectors in a network, given by

$$h^A(t;\theta,\varphi,\psi) = F_+^A(\theta,\varphi,\psi)h_+(t) + F_\times^A(\theta,\varphi,\psi)h_\times(t) \,. \quad (26.21)$$

Here F_+^A and $F_\times^A$ are the *antenna pattern functions* (for a definition see, e.g., [26.15]). (*Schutz* [26.16] calls F_+ and $F_\times$ antenna *amplitude* pattern functions to distinguish them from $F_+^2 + F_\times^2$ that he calls antenna *power* pattern functions.) They are functions of the source position (θ,φ) on the sky and the polarization angle ψ; ψ is the angle between vector $\hat{\theta}$ (which is one of the basis vectors $(\hat{n},\hat{\theta},\hat{\varphi})$ that is naturally defined in a coordinate frame best suited to describe a detector) and the vector $\hat{x}_R$ (which is one of the basis vectors $(\hat{x}_R,\hat{y}_R,\hat{z}_R)$ that is naturally suited to describe the radiation) (Fig. 26.3a). In order to fully reconstruct the incident radiation, a detector must measure five numbers: $h_+, h_\times, \theta, \varphi$, and ψ. A single detector measures one amplitude (namely, the response h^A given above) and the arrival time of an event. It cannot disentangle the two polarizations and the three angles, if the Doppler modulation of the signal due to changing position of the source during the course of observations is unimportant. A long-lived source can, in general, cause modulation of the amplitude and phase of the waves, and these modulations depend on the sky position of the source and also the wave's polarization. Such modulations could allow even a single detector to infer the two polarizations. For transient waves, however, a network of three or more noncollocated detectors would be necessary to measure the various angles and the two polarizations, the only exception being a spherical detector (see below).

Figure 26.3 shows how a network of three detectors could determine the sky position of a source by triangulation. The time delay Δt_k in the arrival time of a signal in a given detector pair defines a circular annulus on the sky for possible source directions, the thickness of the annulus depending on the error in the measurement of time-of-arrival of the signal. Three noncollinear detectors will define, in general, three distinct circles, which intersect at two points. One of these is the true direction $\hat{n}$ to the source and the other $\hat{n}'$ is its mirror reflection in the plane formed by the three detectors [26.17].

Although triangulation does not completely determine the source direction, it turns out that often individual responses h^A, $A = 1, 2, 3$ are only consistent with one of the two degenerate directions [26.18]. Thus, a network of three detectors completely determines the direction to a transient source. In order to obtain the smallest possible error in the sky position, detectors should be as widely spaced as practical, which is equal to the diameter of the earth ≈ 14000 km. The

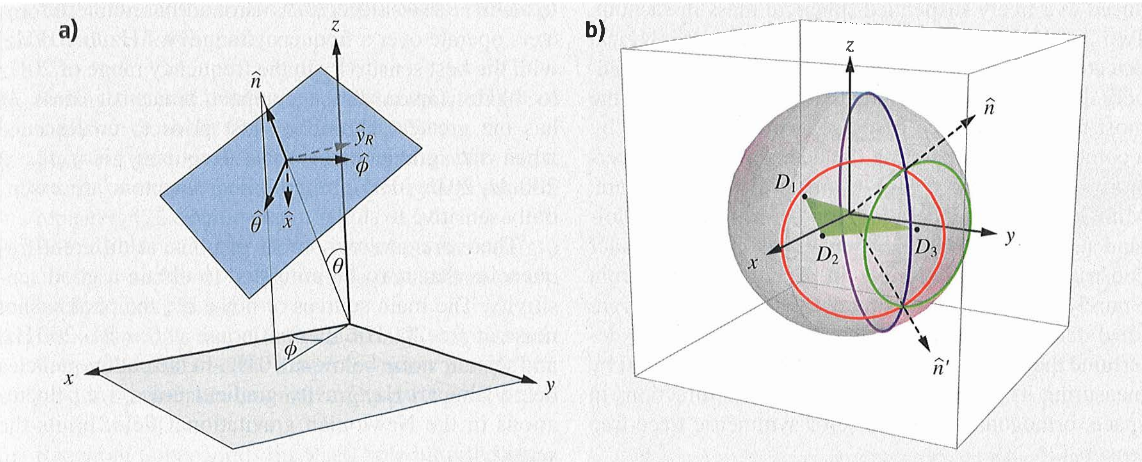


Fig. 26.3 (**a**) The various angles to be determined by a network of detectors. In addition to the sky position (θ, φ) of the source, the polarization angle ψ between the detector preferred coordinate axes $(\hat{\theta}, \hat{\varphi})$ and radiation preferred coordinates $(\hat{x}_R, \hat{y}_R)$ must also be measured. (**b**) A network of detectors D_1, D_2, D_3 can be used to triangulate the sky position (θ, φ) of a gravitational wave source. Arrival times in each detector pair determine only a circular annulus on the sky for possible source directions; *three circles* corresponding to the three detector pairs intersect at two points; one of these is the true direction $\hat{n}$ and the other, $\hat{n}'$, is the mirror image of $\hat{n}$ in the detector plane

angular resolution of a detector network is roughly given by the Rayleigh criterion $\Delta\theta \approx \lambda/D$, where λ is the wavelength of radiation and D is the average baseline of the detector network. For $\lambda = 3000$ km, which corresponds to a frequency of 100 Hz, and $D \approx 10\,000$ km, $\Delta\theta = 0.3$ rad, and so we can expect sources to be identified within a sky patch of ≈ 300 square degrees. This would be smaller by a factor equal to the square of the signal-to-noise ratio (SNR). Therefore, a transient source that produces an SNR of 8 (the minimum SNR required to confidently claim a detection) could be localized within about 5 square degrees by a factor equal to the square of the SNR. Therefore, a transient source that produces an SNR of 8 (the minimum SNR required to confidently claim a detection) could be localized within about 5 square degrees.

26.4.1 Resonant Mass Detectors

In ca. 1959, *Joseph Weber* built the first gravitational wave detector at the University of Maryland, USA [26.19]. His device consisted of a cylindrical aluminium bar, termed a *resonant bar antenna*, 2 m in length and 1 m in diameter. Weber's experiments generated much interest in gravitational wave detectors, and resonant mass detectors were built in Louisiana, Stanford, Rome, CERN, Perugia, Glasgow, Munich, and Perth (for a recent review of bar detectors, see [26.20]).

According to the geodesic deviation equation, a passing gravitational wave would stretch and squeeze the antenna. The induced vibration in the bar is amplified by a transducer and recorded. The main source of noise in resonant antennas is thermal noise and to reduce this as much as possible bar detectors are made of very high-Q, typically $Q \approx 10^6$, and operated at cryogenic temperatures, $T \approx 100$ mK. Bar detectors are narrow band detectors whose operating frequency is $f \approx 500\,\text{Hz}-1.5\,\text{kHz}$ and so the root-mean-square amplitude of thermal vibrations of a bar of mass $M = 10^3$ kg is $\sqrt{\langle\delta\ell^2\rangle} = \sqrt{kT/(4\pi^2 f^2 MQ)} \simeq 6\times10^{-21}$ m. Since $\ell \approx 2$ m, a wave of amplitude h causes vibrations of amplitude $\delta\ell = h\ell/2 = h$ m, the sensitivity of a bar will be limited to waves of amplitude $h = 10\times\sqrt{\langle\delta\ell^2\rangle} = 6\times10^{-20}$, where a factor of 10 is included to account for the SNR needed for a detection. Most bar detectors are narrow-band detectors, with a bandwidth of about 20–50 Hz near a kHz. They are sensitive to strain amplitudes $h \gtrsim 10^{-20}$ if the source radiates in this frequency window.

Another type of resonant mass detector that is very attractive is a spherical antenna. The basic idea here is the same as that of a bar detector, namely to detect gravitational waves by monitoring the oscillations in-

duced in a freely suspended spherical mass in vacuum. Two significant efforts to build and operate spherical detectors are the MiniGrail [26.21] and *Mario Schenberg* [26.22] projects. Here too thermal noise is the most dominant source of noise, which is reduced by a combination of operating the detector at low temperature and choosing a high-Q material for the sphere. MiniGrail was in operation for a few years in Holland [26.21] and Mario Schenberg is currently under construction/commissioning in Brazil [26.22]. From a purely theoretical point of view spherical antennas are ideal detectors. A single detector can, in principle, determine the sky position of a source. This is achieved by measuring its response in five *orthogonal* directions in space, orthogonal in the sense of symmetric trace-free tensors [26.23].

26.4.2 Laser Interferometers

Resonant mass detectors operate at a frequency of $\approx$ kHz and their sensitivity bandwidth is typically 20–50 Hz at best. Many astronomical sources are expected to emit at far lower frequencies and one of the most important sources, namely coalescing compact binaries, produces a wide band signal. The chance of detecting gravitational waves is far greater with a detector that operates at lower frequencies and has a wide bandwidth. Soon after Weber began operating his bar detector, and in ca. 1962 *Gertsenshtein* and *Pustovoit* conceived the idea of an interferometric gravitational wave detector [26.24]. The basic idea of an interferometric gravitational wave detector is to compare the time of flight of laser beams in the two arms of a Michelson interferometer to infer the presence of gravitational waves. Many such detectors have been built and operated on the ground and a number of proposals have been made to develop space missions. We will discuss these in turn.

Ground-Based Detectors. Ground-based interferometers operate over a frequency range of 1 Hz to 10 kHz, with the best sensitivity in the frequency range of 20Hz to 2 kHz. Given that a compact binary of mass M has the greatest luminosity just prior to coalescence when the gravitational wave frequency is $f_{\text{merge}} \approx 200(M/20M_{\odot})$Hz, ground-based detectors are essentially sensitive to stellar mass sources.

There are several sources of noise at different frequencies that must be mitigated to obtain a good sensitivity. The main sources of noise are the photon shot noise at $f > 200$Hz, thermal noise at $f \approx 50$–200 Hz, and seismic noise below ≈ 50 Hz. In fact, at frequencies below about 1 Hz, gravity gradient noise, i. e., fluctuations in the Newtonian gravitational field, limits the sensitivity.

Interferometric detectors began with prototypes at MIT, Caltech, Glasgow, and Munich (for a historical review, see [26.25]). Coincident operation of the Glasgow and Munich prototypes for 100 h [26.26] was one of the earliest demonstrations that the technology to build and operate long-baseline detectors was becoming available. The Caltech 40 m prototype led to the American Laser Interferometer Gravitational-Wave Observatory (LIGO) [26.27]. LIGO consists of three interferometers at two sites: a 4 km interferometer at Livingston, LA (Fig. 26.4a) and one 4 km and one 2 km interferometers at Hanford, WA (Fig. 26.4b). The French-Italian Virgo [26.28] is a 3 km baseline interferometer in Pisa, Italy (Fig. 26.4c), with sensitivity comparable to LIGO. The Japanese TAMA [26.29] and the British-German GEO600 [26.30] are 300 m and 600 m interferometers, respectively, with sensitivity levels considerably smaller than LIGO and Virgo.

These detectors operated between ca. 2003–2010 and took data over several science runs. Although they did not make any detections, they not only reached very impressive sensitivity levels [26.31] (Fig. 26.5) but also

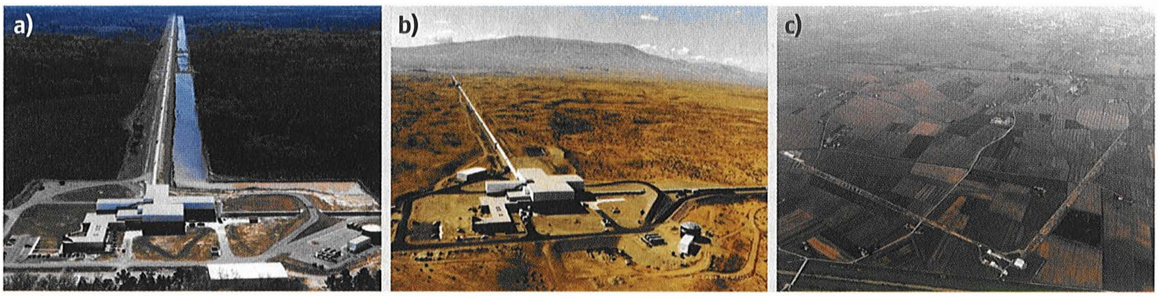


Fig. 26.4a–c Aerial view of (**a**) LIGO Livingston, (**b**) LIGO Hanford, and (**c**) Virgo detectors

reached important astrophysical milestones. These include, but are not limited to:

1. Setting the best ever upper limit on the strength of radiation from the Crab pulsar [26.32].
2. Showing that certain short hard gamma-ray burst events of extragalactic origin might be soft gamma repeaters [26.33]
3. Beating the nucleosynthesis bound on the strength of the stochastic primordial gravitational wave background [26.34].
4. Demonstrating that ellipticity of many known millisecond pulsars is less than a few parts per million (and in some cases less than a few parts per 10 million) [26.32].
5. Reaching upper limits on black hole binary coalescence rates [26.35] close to astrophysical predictions [26.36]
6. A search for gravitational waves in coincidence with 154 GRB [26.37] that occurred during the final data taking of the LIGO and Virgo detectors before they were shutdown for upgrades.

LIGO and Virgo detectors are being upgraded with target strain sensitivities for LIGO as shown in Fig. 26.5 [26.38]. They are expected to reach these sensitivity levels on the time scale of the next 2–5 yr. Additionally, Japan is building a new underground detector called KAGRA, which will eventually deploy cryogenic mirrors to beat the thermal noise – a key technology for future detectors.

A global network of advanced detectors with sensitivity levels good enough to make first direct detection of gravitational waves is expected to be operational before the end of this decade. However, new ideas are already being pursued in order to improve the baseline and science return of networks. The LIGO project is planning to move one of the two detectors at the Hanford observatory to India [26.39]. A detector in India, currently under consideration by Indian funding agencies, will help improve source localization, break the degeneracy between different physical parameters (most importantly the distance to a binary source and the inclination angle of the binary), and also improve the lifetime of the network [26.16, 39, 40].

Figure 26.6 shows how a network of four gravitational wave detectors LIGO-Hanford, LIGO-India, LIGO-Livingston, and Virgo (HILV) can have quite good sky coverage and achieve good sky localization

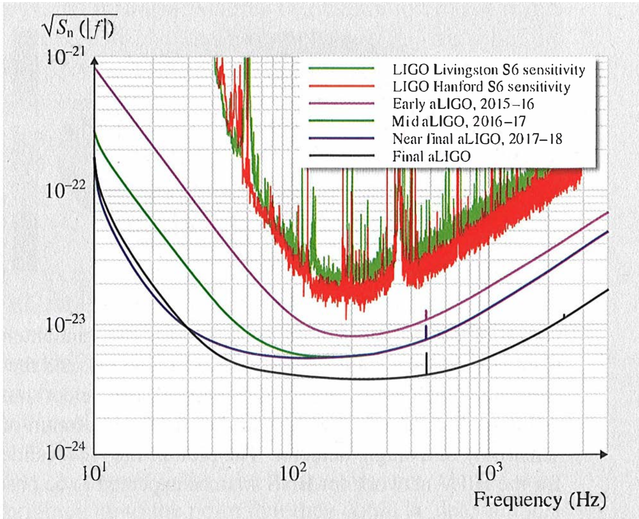


Fig. 26.5 Sensitivity of LIGO detectors during the sixth science run (*top two curves*) and expected sensitivity at various stages during the commissioning of advanced LIGO (aLIGO). Both the sensitivity level at various stages and the schedule of aLIGO are tentative and subject to change. (aLIGO sensitivity curves are from [26.38], sensitivity curves for the sixth science run of initial LIGO (iLIGO) are from [26.41])

of binary inspiral sources. The diagram on the left-hand side plots the quantity

$$F^2 \equiv \sum_{A=\mathrm{H,I,L,V}} (F_+^A)^2 + (F_\times^A)^2 .$$

This is a measure of the sensitivity of the network to different parts of the sky. A single detector has a quadrupole antenna pattern that has directions along which the detector is completely blind. A network of antennas, however, will be sensitive to almost the entire sky, although its sensitivity will not be isotropic. The diagram on the right-hand side shows the angular resolution of the HILV network for binary neutron stars at a distance of 180 Mpc. The angular resolution tends to be better along directions of good sky sensitivity with about 50% of the sources resolved to within 20 square degrees [26.40].

Einstein Telescope. In Europe, a conceptual design study to build an underground detector, called Einstein telescope (ET), has just been completed [26.43, 44]. It will have a triangular topology with 10 km arms and operate three broadband detectors at a single site by using

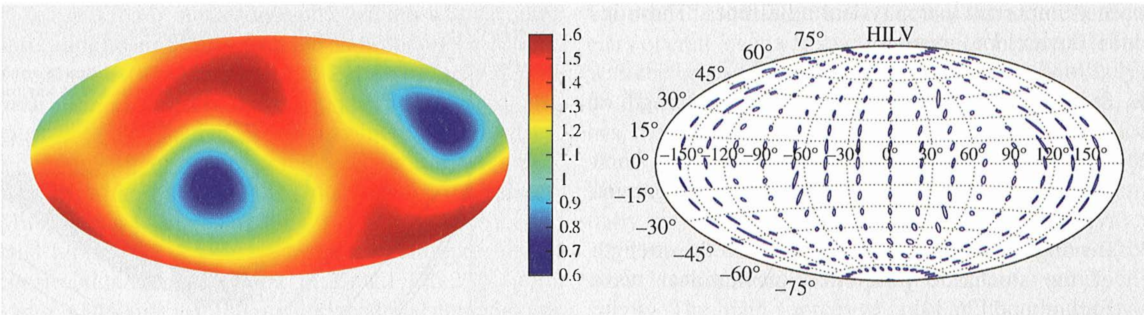


Fig. 26.6 Mollweide plot *on the left* shows the joint antenna pattern of the HILV network, i. e., a network consisting of LIGO detectors at Hanford, India, and Livingston, and the Virgo detector in Pisa, averaged over the polarization angle ψ and source inclination angle ι. The network can detect sources almost anywhere in the sky but has different sensitivity to different directions in the sky, varying from a minimum of 0.6–1.6, where a sensitivity of 1 corresponds to the best sky sensitivity of a single detector. The plot *on the right* (slide courtesy *S. Fairhurst* [26.42]) gives error ellipses on the sky for the HILV network for BNS sources expected to be observed in the advanced detector network

each arm of the triangle twice. With a strain sensitivity that is 10 times better than that of advanced detectors, ET should be able to take a census of stellar mass BBH up to a redshift $z \approx 17$–20, detect intermediate mass black hole binaries at redshifts of $z \approx 5$–7, and BNS at $z \approx 2$–4 [26.45].

Laser Interferometer Space Antenna. Sensitivity of ground-based detectors below a few Hz will be limited by gravity gradient noise that arises as a result of variations in the surface density of Earth due to seismic waves, variations in the density of air caused by wind and other environmental factors and, more generally, anthropogenic noise [26.46]. Some of these noise sources can be reduced by building a deep underground detector (as KAGRA have done and ET is planned) where density of air and anthropogenic noise will cease to be problems and the effect of seismic waves greatly suppressed.

Another solution to low-frequency noise sources is to place a detector in space. The Laser Interferometer Space Antenna (LISA) in Europe and the US [26.47] and DECIGO in Japan [26.48] are two projects that aim to have free flying spacecraft in heliocentric orbit, away from the Earth. For example, LISA constitutes a set of three spacecraft, separated from each other by 5 million km, flying in a triangular formation in heliocentric orbit. LISA will be sensitive to sources in the frequency interval of [0.01, 100] mHz. Space antennas like LISA can probe radiation from supermassive black hole binaries from far corners of the Universe as well as galactic white dwarf binaries [26.15].

26.4.3 Pulsar Timing Arrays

A population of highly stable millisecond pulsars, with timing accuracies of ≈ 100 ns over several years, could serve as an array of clocks whose regular ticks would be coherently modulated due to gravitational waves passing by the Earth. There is worldwide effort to observe stable millisecond pulsars and exploit them for detecting gravitational waves, so-called *pulsar timing arrays* (PTAs) [26.49–53]. Precise timing of an array of pulsars, may detect nanohertz gravitational radiation that one might expect from *merging* supermassive black hole binaries of masses in the range $[10^9, 10^{10}]M_\odot$, but they may also be sensitive to binaries of lower masses at an earlier stage in their evolution [26.54, 55]. More importantly, the array will also be sensitive to stochastic gravitational waves of nanohertz frequencies [26.56]. Indeed, the current best constraints on primordial gravitational wave background are obtained by PTAs [26.57] at frequencies of $\approx 10^{-9}$ Hz to be $\Omega_{GW} < 2 \times 10^{-8}$, where Ω_{GW} denotes the energy density in gravitational waves compared to the closure density of the Universe.

Stochastic gravitational waves of still lower frequencies (wavelengths as large as the Hubble radius of the Universe) might systematically affect the polarization patterns of the cosmic microwave background [26.58]. So far, only upper limits have been set by CMB experiments, but the Planck satellite could either detect or set the most stringent limits on the strength of primordial gravitational radiation produced in the inflationary era [26.59].

In summary, there are many opportunities to directly observe gravitational radiation, and it is widely expected that this will happen before the end of the current decade. Gravitational wave observations should help answer many puzzles in astronomy and cosmology, but this new window of observation might reveal sources and phenomena no one has imagined before. It is the quest for the unknown that makes the field so exciting.

26.5 Gravitational Astronomy

Astronomical sources with the greatest compactness are the most luminous sources of gravitational waves. Therefore, the brightest sources are neutron stars and black holes – the most compact objects in the Universe. Radiation back reaction determines the dynamics of most luminous systems, either driving them to instability that makes them catastrophically bright (e.g., compact binary star coalescences) or causing them to shut down emission by decreasing the nonspherical motion in the system (e.g., rotating neutron stars).

In this section we will discuss sources of gravitational radiation, their strengths and how often we expect to see transient sources. Figure 26.7b plots the strengths of a number of potential astronomical sources together with the sensitivity of some of the detectors that have either already been built and are being operated (iLIGO) or currently under construction (aLIGO), and others that are planned for the future (ET, LISA/eLISA). For the sake of clarity we have not shown the sensitivities of initial and advanced Virgo that are similar to iLIGO and aLIGO, respectively, but with a slightly better low-frequency sensitivity, KAGRA, which will be similar to aLIGO, and GEO600, whose sensitivity above 800 Hz will be about a factor ≈ 10 worse than aLIGO but significantly poorer sensitivity at lower frequencies. Figure 26.7a shows sources expected to be seen in the millihertz frequency region and sensitivity of LISA. Here we have left out the sensitivity curve of DECIGO, designed to operate in the frequency window of 100 mHz to 20 Hz with a strain sensitivity of $5\times10^{-26}\,\mathrm{Hz}^{-1/2}$.

26.5.1 Compact Binaries

Binaries consisting of a pair of neutron stars, a pair of black holes, or a neutron star and a black hole are called compact binaries. As far as we know, they are the most powerful emitters of gravitational radiation. The Hulse–Taylor binary pulsar is a typical example of such a system that consists of a pair of $\approx 1.4M_\odot$ neutron stars. With an orbital period of 7.5 h and eccentricity of 0.62, this system loses energy to gravitational waves at the rate of $\approx 7\times10^{24}\,\mathrm{J\,s^{-1}}$. It will be another 300 million years for gravitational radiation to drive this system to coalescence. Its cousin, J0737-3039 discovered in 2003 [26.60], is a double pulsar (the only known such system) whose orbital period is only 2.5 h and it will coalesce only in 85 million years.

These systems spend millions of years with very low luminosity. As the two stars get closer, they brighten up to the point that they could be detected just minutes to seconds prior to merger even at cosmological distances. Their late time dynamics is believed to be governed entirely by gravitational radiation back reaction, which causes the eccentricity of these systems to become negligible well before they become observable by ground-based detectors.

Using post-Newtonian (PN) approximation methods [26.10], it has been possible to calculate the waveforms from these systems very accurately. At the lowest order approximation (i. e., using the quadrupole formula given in (26.9) and (26.11)) the plus and cross polarizations of the waves for a binary, whose distance from the Earth is r, orbital orientation with respect to the line of sight is ι, and orbital separation is R, take the form

$$\begin{aligned} h_+(t) &= \frac{2\mathcal{M}}{r}(\mathcal{M}\omega)^{2/3}(1+\cos^2\iota)\cos 2\phi\,,\\ h_\times(t) &= \frac{4\mathcal{M}}{r}(\mathcal{M}\omega)^{2/3}\cos\iota\sin 2\phi\,. \end{aligned} \tag{26.22}$$

Here $\mathcal{M}=\nu^{3/5}M$ is called the *chirp mass*, $\nu = m_1m_2/M^2$ is the symmetric mass ratio, $\omega = \mathrm{d}\phi/\mathrm{d}t = \sqrt{GM/R^3}$ is the angular frequency, and ϕ is the orbital phase. One can solve for $\omega(t)$ and $\phi(t)$ using the energy balance equation $\mathcal{L} = -\mathrm{d}E/\mathrm{d}t$, which states that the luminosity in gravitational waves is generated by the loss in energy E of the system. Noting that $E=-\nu M^2/2R$ and $\mathcal{L}=(32/5)\nu^2(M\omega)^{10/3}$, we obtain

$$\begin{aligned} \dot\omega &= \frac{96}{5}\mathcal{M}^{5/3}\omega^{11/3} \quad\Rightarrow\quad \omega(t) = \omega_\mathrm{i}\left(1-\frac{t}{t_\mathrm{C}}\right)^{-3/8},\\ \phi(t) &= -\frac{1}{32}(\mathcal{M}\omega)^{-5/3}\,. \end{aligned} \tag{26.23}$$

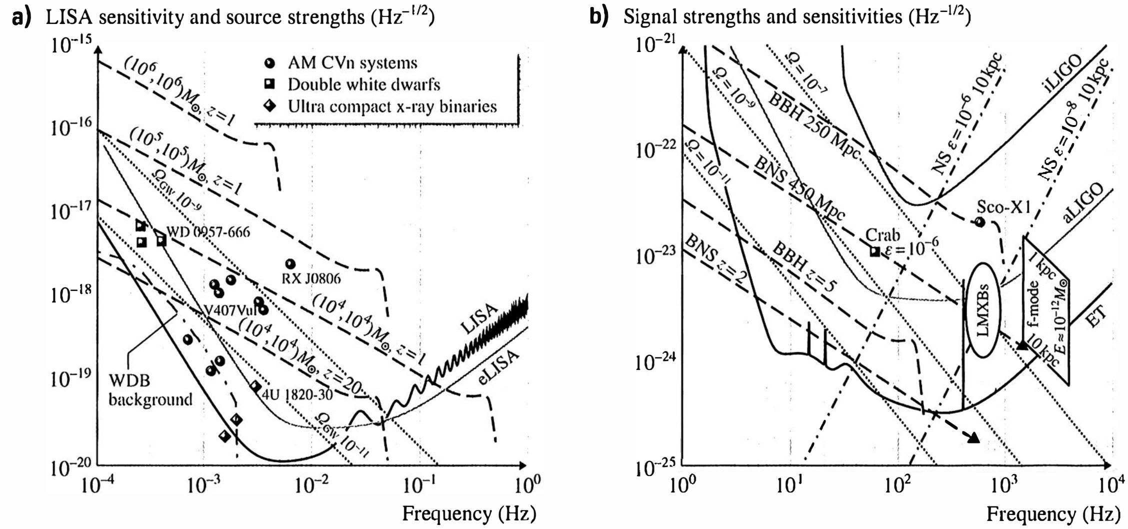


Fig. 26.7 (**a**) Sensitivity of LISA and its variant termed eLISA. (**b**) Sensitivity of three generations of ground-based detectors, iLIGO, aLIGO, and ET. Inspiral sources are assumed to be randomly oriented and located with respect to the detectors. Binary neutron stars (BNS) are assumed to consist of two $1.4M_\odot$ neutron stars; only the inspiral part of the signal is shown up to the last stable orbit when the frequency of the source is $f_{\mathrm{LSO}} = c^3/(6^{3/2}G\pi M)$. The merger part of BNS signal could be very complicated and is not fully understood. Binary black holes (BBH) are assumed to consist of two $10M_\odot$ black holes; here the inspiral phase smoothly transitions to merger phase, followed by quasi-normal mode ringing of the final black hole. Both (**a**) and (**b**) show the characteristic amplitude h_c (in units of $\mathrm{Hz}^{-1/2}$) for a number of sources: for burst sources of Fourier amplitude $H(f)$ the characteristic amplitude is $h_c \equiv 2\sqrt{f}H(f)$; for continuous wave sources of strain amplitude h_0 the characteristic amplitude after integrating for a time T is $h_c \equiv \sqrt{T}h_0$; and for a stochastic background of spectral density $S_h(f)$ the characteristic amplitude at frequency f after cross-correlating data from a pair of detectors over time T is $h_c \equiv (Tf)^{1/4}\sqrt{S_h(f)}$. For continuous waves and stochastic radiation, we assume the period of integration to be $T = 1$ yr. The sensitivity and source strengths are all in units of $\mathrm{Hz}^{-1/2}$

Here ω_i is the angular frequency at the orbital separation R_i. It is immediately clear that the amplitude and frequency of the emitted waves increase with time, producing a characteristic chirp signal. Since the phase and amplitude evolutions are known accurately it is possible to dig these signals out of background noise using matched filtering.

Figure 26.8 is a frequency-mass diagram in which we show the frequency range accessible to various detectors and some interesting features deduced from the above equations:

1. The frequency at which the two stars merge depends on the total mass and it is roughly given by $f_{\mathrm{merge}} = c^3/(6^{3/2}\pi GM) \simeq (M/10M_\odot)^{-1}440\,\mathrm{Hz}$. Figure 26.8 shows this frequency as a function of the binary mass as a dot-dashed line marked LSO. The peak luminosity of a binary reaches soon after the system reaches this frequency.
2. The merged object, a black hole, emits quasi-normal modes (QNM) to get rid of the deformation inherited in the process of merger. The frequency of the fundamental mode $\approx (M/M_\odot)^{-1}1.2\,\mathrm{kHz}$, is shown as a function of mass as a dashed line immediately above the LSO line. Although higher-order modes might be excited it is not expected that they will carry too much energy and so one does not expect any radiation to occur in nature in the grey shaded region above this line.
3. A circular binary starting from frequency f_i coalesces roughly on a timescale t_C given by (26.10), where $\kappa_i = (\pi GMf_i)^{2/3}/c^2$. Dotted lines in Fig. 26.8 show the starting frequency, as a function of total mass for equal mass binaries (i.e., $\nu = 1/4$), from where the system would last for 1 min, 1 d, and 1 yr.
4. The rate at which the frequency of gravitational waves changes $\dot{f} = \dot{\omega}/\pi$ (cf. (26.23)) depends on

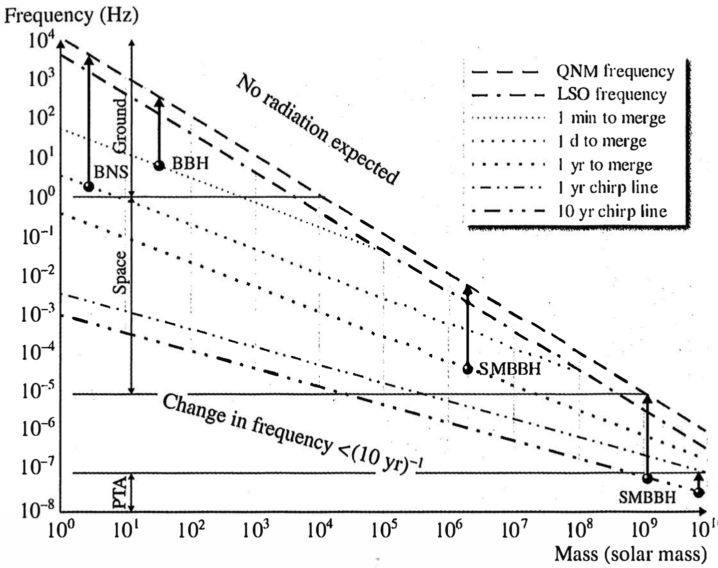


Fig. 26.8 Frequency–mass plot showing typical compact binary sources expected to be observed in ground-based detectors (LIGO, Virgo, KAGRA, ET), space-based detectors (LISA), and PTA. See text for details

the chirp mass of the system. For binaries whose frequency changes during the course of observation it will be possible to determine the chirp mass of the system. If the system is observed for a time T, then the smallest change observable is $1/T$ and the change in frequency is $T\dot{f}$. Equating the two gives the observation time necessary to observe the chirping of a binary, namely, $T = (\dot{f})^{-1/2}$. Chirp mass will be measurable over a period of 1 (10) year for those systems that are on the 1 yr *chirp line* (10 yr *chirp line*), (dot-dot-dashed lines). It is not possible to measure any physical parameters for systems below this line if the observation period is $\lesssim 10$ yr.

As mentioned earlier, (26.22) and (26.23) are derived in the quadrupole approximation. There are PN corrections, currently known to order $(v/c)^7$ beyond the lowest order, that have additional dependences on the mass ratio of the system. Moreover, if the binary is on an eccentric orbit or the component stars have large spins, the signal's frequency and amplitude will have modulations. Note also that the ratio of the two polarizations contains important information about the orientation of the source (namely, the angle ι). Thus, imprinted in the structure of the signal are the various parameters of the source. By a careful analysis it will be possible, in principle, to measure the masses of the component stars, eccentricity of the orbit, spins of the two bodies, and the polarization of the radiation. In reality, however, not all parameters of the source can be measured to a good accuracy as there are strong correlations between some of the parameters (for example, ι and r are strongly correlated) [26.18].

The PN approximation breaks down when the two objects are very close. In the case of BNS it is very hard to model the merger phase of the evolution that involves strong tidal interaction of the two bodies. Observing this phase could lead to very useful insights into the structure of matter under extreme conditions of density, temperature, pressure, and magnetic fields. This phase, however, is not likely to be visible in advanced detectors but could be observed in ET. In the case of BBH, analytical methods [26.61, 62] and numerical relativity simulations [26.63] have been successfully employed to understand the merger phase. The slow adiabatic phase smoothly transitions to a brief merger phase when the luminosity reaches its peak, followed by the quasi-normal mode ringing of the final deformed black hole (Fig. 26.7).

Matched Filtering and Signal Visibility. Since compact binary signals are well-modeled it is possible to dig out signals buried in noise using matched filtering [26.64]. The SNR ρ obtained by cross correlating the data with an optimal template is given by [26.65]

$$\rho = \left[4 \int_{f_{\min}}^{f_{\max}} \frac{|H(f)|^2}{S_h(f)} \, \mathrm{d}f \right]^{1/2} ,$$

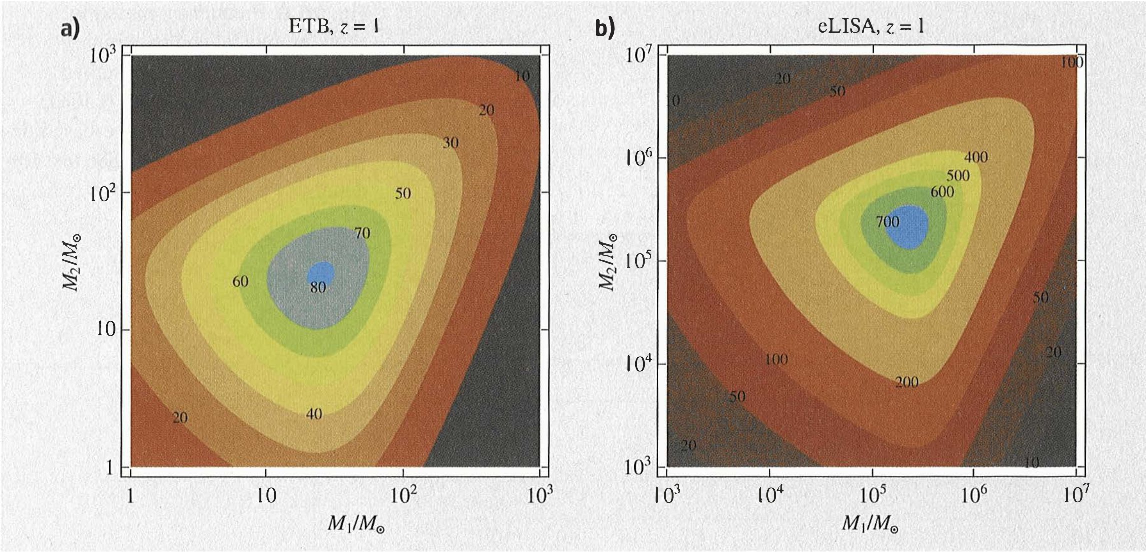


Fig.26.9a,b Contour map of optimal SNR in the inspiral phase of compact binary sources in **(a)** ET and **(b)** eLISA

where $H(f)$ is the Fourier transform of the signal $h(t)$ and $S_h(f)$ is the noise spectral density of the detector in question. $f_{\min}$ and $f_{\max}$ are the appropriately chosen lower and upper frequency cutoffs used in computing the SNR. For example, in the case of ground-based detectors the upper limit is the frequency at which the binary coalesces, which is taken to be the last stable orbit frequency for BNS $f_{\mathrm{LSO}} = c^3/(6^{3/2}\pi GM)$ or the quasi-normal mode frequency for BBH $f_{\mathrm{QNM}} = 1.2(M/10M_\odot)^{-1}$kHz and the lower limit is taken to be 10 Hz for advanced detectors and 1 Hz for ET. In the case of eLISA and LISA one could choose $f_{\max}$ as in the case of ground-based detectors but one should be careful not to choose $f_{\min}$ to be too low, as binaries could spend millions of years in the sensitivity band of eLISA. One typically takes the lower cutoff such that it takes ≈ 3 yr for the binary to coalesce starting from $f_{\min}$.

Figure 26.9 shows the contour map of the optimal signal-to-noise ratio (i. e., SNR assuming optimal sky position and orientation of the binary) for binaries consisting of masses M_1 and M_2 and at a redshift of $z = 1$. The upper and lower frequency cutoffs are chosen as discussed above. Quite clearly, both of these instruments will observe sources at $z \gg 1$ and could provide clues as to the birth and evolution of first black holes and their demographics. Note also that eLISA and ET observe different ranges of masses and, therefore, complement each other in exploring black holes at cosmological distances.

Stellar Mass Compact Binaries. The merger time of Hulse–Taylor and J0707-3039 binaries is far less than the Hubble time. We can, therefore, expect that the BNS merger is not an uncommon event in the Universe. How many such events might we expect each year within a given volume of the Universe? Unfortunately, the small number of observed galactic systems is only able to provide an estimate that is uncertain by a factor of $\approx 10^3$. Nominally, the rate could be about one event within a volume of 400 Mpc^3 but the rate could be lower by a factor of 100 or larger by a factor of 10 [26.66].

The other two categories of stellar mass binaries, a pair of black holes (BBH) or a neutron star and a black hole (NSBH), have so far not been observed, although *Belczynski* et al. argue that high mass x-ray binaries IC10 X-1 and NGC300 X-1 are progenitors of BBH [26.67]. If this is true then statistical arguments similar to the one applied to BNS give the event rate of BBH to be $R = 3.36^{+8.29}_{-2.92}$ in initial LIGO and roughly 1000 greater in aLIGO [26.67]. Some authors [26.68] have explored the effect of metallicity on the formation and evolution of massive stars to deduce that black hole mergers could be far more common in the Universe with rates in excess of several 100s to several 1000s in aLIGO.

In all cases the rates are rather uncertain. However, a network of advanced gravitational wave detectors could have a distance reach of 200 Mpc for BNS, 800 Mpc for NSBH, and 1.5 Gpc for BBH, within

which we can not only expect to make the first direct detection of gravitational waves, but also place a firm constraint on astrophysical models of the formation and evolution of compact stars and their merger rates.

Supermassive and Intermediate Mass Black Hole Binaries. There is now strong observational evidence that galactic centers host supermassive black holes, i. e., black holes of millions to several tens of billions of solar masses. Decades of observations of stellar orbits close to the galactic center have revealed the presence of a black hole of $4 \times 10^6 M_\odot$ in the nucleus of our own Milky Way [26.69, 70]. When and how did such black holes form? Did the black holes precede the galaxies or did they form after the galaxies were assembled? What were their initial masses and how did they grow? These are among the most pressing unsolved questions in cosmology. Gravitational wave observations in different spectral windows might be able to answer some of these questions.

A binary of intrinsic masses $(10^4, 10^4)M_\odot$ at $z = 20$ will appear to us as a $(2.1 \times 10^5, 2.1 \times 10^5)M_\odot$ binary and merge at a frequency of around 10 mHz. The redshift effect on the amplitude is so great that a binary that is visible with a modest SNR (of, say, 20) at redshift $z = 1$ will continue to be visible until the observed mass is so large that the binary merges outside the sensitivity band of the detector (see Fig. 26.7). Even at a redshift of $z = 20$ a randomly oriented $(10^4, 10^4)M_\odot$ binary with a random sky position would produce, close to merger, a characteristic amplitude of $h_c \equiv 2\sqrt{f}|H(f)| \approx 1.4 \times 10^{-19}\,\mathrm{Hz}^{-1/2}$, and will be clearly visible in LISA/eLISA. These are such high redshifts that the Universe was probably assembling its first black holes at this epoch. LISA/eLISA can take a census of supermassive black hole binaries in the mass range 10^4–$10^7 M_\odot$ in the entire Universe and provide the necessary input for testing different scenarios of the formation and growth of galaxies [26.71, 72].

The merger rate of supermassive black holes is highly uncertain as there are only a handful of such candidate binaries that would merge within the Hubble time. Detailed modeling of these systems is very difficult due to many unknown astrophysical parameters, including their masses and spins when they formed and how they grew. Predicted merger rates in the Universe in the range of masses that LISA/eLISA could observe are of order ≈ 30–$100\,\mathrm{yr}^{-1}$, depending on the model used for the formation and growth of massive black holes, of which ≈ 20–30 should be detectable by LISA [26.73].

There is as yet no conclusive evidence for the existence of black holes of mass in the range $\approx 10^2$–$10^4 M_\odot$, the so-called *intermediate mass* black holes. However, there are strong indications that certain ultra-luminous x-ray sources, e.g., HLX-1 in ESO 243-49 [26.74], are host to intermediate mass black holes. If a population of such binaries exists and they grow by merger, then, depending on their masses, ET will be able to explore their dynamics out to $z \approx 6$ and study their mass function, redshift distribution, and evolution. Several authors have looked at the possibility that intermediate mass black hole binaries may form and merge in dense stellar clusters. Modeling the growth of seed black holes using different scenarios, these authors conclude that ET could observe a few to a few tens of intermediate mass black hole binaries per year [26.75–77].

Extreme Mass Ratio Binaries. The binaries that we have discussed so far have component stars of comparable masses. When one of the masses is far smaller than its companion, we have the problem of a test body in *near* geodesic motion in black hole geometry. Such binaries are called *extreme mass ratio binaries,* as the mass ratio m_1/m_2, $m_1 \gg m_2$, could get stupendously large. The orbits in this case are near geodesics because in each cycle the test body loses only a small amount of its rotational energy to radiation; recall that the luminosity of a binary source is $\nu^2 \simeq (m_2/m_1)^2$.

In general relativity, bound geodesics in Kerr spacetime geometry have a far richer structure than the simple elliptic orbits of Newtonian gravity. Geodesics with eccentricity close to unity and periastron very near the horizon are especially intriguing. A test particle on such an orbit could begin at apastron and when it reaches periastron it might exhibit tens of near-circular orbits, before ending up at an entirely different point in space as apastron. When the particle is close to periastron, it loses far more energy (remember that the luminosity is proportional to fifth power of compactness) than when it is at apastron and hence the emitted gravitational wave signal can have a rather complex structure.

Moreover, such geodesics could be truly spherical orbits and sample the entire spacetime region near a black hole. Imprinted in the signal's phasing is the full multipole structure of the black hole spacetime and it should be possible to produce a map of the Kerr geometry of the central object by observing the inspiral of an extreme mass ratio binary [26.78]. This would be a stringent test of general relativity as it would be pos-

sible to check if the spacetime geometry of black holes is truly described by only their masses and spins or if black holes have *hair*.

LISA is best suited to observe extreme mass ratio inspirals. Supermassive black holes at galactic nuclei are believed to grow by the infall of stellar mass and intermediate mass black holes. Such events could be observed by LISA at cosmological distances. For instance, the inspiral of a $10M_\odot$ black hole into a $10^6 M_\odot$ supermassive black hole at $z = 1$ would produce a detectable amplitude in LISA. The rates in this case are also highly uncertain and range from a few to several hundreds per year [26.75, 79].

Gamma-Ray Bursts. Gamma-ray bursts (GRBs) are extremely bright flashes of energy that last anywhere from milliseconds to several minutes. Discovered in the late 1960s by US spy satellites, they are the most luminous known sources in the Universe. Their rapid variability over short time scales implies that they are very compact sources, likely neutron stars or stellar mass black holes. By measuring the redshift of host galaxies it is now known that most GRBs are cosmological in origin. The flux levels often exceed $10^{-8}\,\mathrm{J\,m^{-2}\,s^{-1}}$, implying an isotropic luminosity of $10^{44}\,\mathrm{J\,s^{-1}}$ for bursts at 1 Gpc. This is several orders of magnitude larger than the luminosity of an entire galaxy at all wavelengths. The difficulty in modeling these sources is that it is impossible to produce such stupendously large luminosities from highly compact objects. If the emission is beamed in a narrow cone, however, then the energy requirements can be considerably smaller. Most models assume that the radiation is confined to a cone with an opening angle of about 20°.

If the GRBs are compact sources, then it is plausible that gamma ray emission is accompanied (most likely, preceded) by the emission of gravitational waves. GRBs are classified based on the duration of bursts and spectral hardness. Bursts that last for 2 s or more and with soft spectra are called *long GRBs,* and they are associated with core-collapse supernovae that are expected to emit a burst of gravitational waves before the creation of the fireball leading to GRBs. Bursts lasting for shorter periods of 2 s or less and with hard spectra are termed *short GRBs*, and the most popular progenitor model for such bursts is the coalescence of BNS, which are also the most promising sources for interferometric gravitational wave detectors. It is thought that some of these short GRBs are giant magnetar flares which could also be accompanied by the emission of gravitational radiation.

Observing gravitational waves in coincidence with GRBs will have a tremendous impact on the understanding the progenitors of GRBs and how they are powered. Initial detectors have already placed some impressive constraints on nearby GRBs and set upper limits on the strength of gravitational waves from a population of bursts that occurred during recent science runs [26.33]. Moreover, such coincident observations will help identify the host galaxy and measure its redshift. If progenitors of short GRBs are BNS mergers then this would help measure both the luminosity distance and redshift to the source, *without* the use of the cosmic distance ladder [26.80]. Clearly, such observations will have a great potential for precision cosmography [26.81–84].

GRBs are detected roughly twice a day, of which about a quarter (≈ 170) are short GRBs and the rest (≈ 500) are long GRBs [26.85]. The local rate, inferred from redshift measurements of the host galaxies of a subset of the population, for short GRBs is $10\,\mathrm{Gpc^{-3}\,yr^{-1}}$, an order of magnitude more than that for long GRBs, $0.5\,\mathrm{Gpc^{-3}\,yr^{-1}}$. Most of these events are at redshifts that are not accessible to advanced detectors. Long GRBs, associated with core collapse supernovae, will be particularly faint in the gravitational window at distances greater than about a few Mpc. Short GRBs, associated with BNS or neutron star-black hole binaries, could be observed up to distances in the range of 400–800 Mpc, depending on the total mass of the binary. Advanced detectors might observe coincidences with short GRBs within about 16 months of observation at modest sensitivities [26.84]. ET, however, will be sensitive to neutron star coalescences at $z \approx 2$–4 and will observe them in coincidence with GRBs far more frequently [26.82].

Standard Sirens. Compact binary sources are quite unique for astrophysics and cosmology as they are *standard candles* or, perhaps more appropriately, *standard sirens.* Gravitational wave observations can measure both ω and $\dot{\omega}$ if the observation time T is sufficiently long, namely $T > (\dot{\omega}/2\pi)^{-1/2}$. For example, the change in frequency of a $10^6 M_\odot$ supermassive black hole binary with an orbital frequency of $5\,\mu\mathrm{Hz}$ will not be observable even after 1 year, while that of a $2.8M_\odot$ neutron star binary at 1 Hz would be detectable after roughly 5 min. Systems whose frequencies change during the course of observation are called *chirping* binaries. For binaries that chirp, it is possible to measure the chirp mass of the binary, but since we can also measure the two polarizations h_+ and $h_\times$, we can deduce

the distance D to the binary from gravitational wave observations alone [26.80].

Until recently, it was thought that while gravitational wave observations of a compact binary inspiral can be used to infer the luminosity distance of its host, it will not be possible to measure the host's redshift. This is because from the expression for $\dot{\omega}$ in (26.23) we see that cosmological redshift in frequency ($\omega \to \omega/(1+z)$, $\dot{\omega} \to \dot{\omega}/(1+z)^2$), causes the source to appear to have a larger chirp mass ($\mathcal{M} \to (1+z)\mathcal{M}$). There is no way to tell if the source is at a lower redshift with an intrinsically larger chirp mass or at a higher redshift with an intrinsically smaller chirp mass. It was, therefore, thought that the only way to break the redshift-mass degeneracy is to identify the host galaxy and measure its redshift. While this is still true for BBH, for binaries in which at least one of them is a neutron star it might be possible to infer the intrinsic mass of the system from gravitational wave observations. This is because the equation of state dependent tidal effects, which appear at the fifth PN order, depend on the density M/R^3 of the neutron star and not just on the compactness. It turns out that the tidal effect can be used to determine the source's redshift provided the neutron star equation of state is known [26.86]. The effect is completely absent for black hole binaries, weaker in the case of neutron star-black hole binaries, and quickly diminishes as the mass ratio ν decreases.

26.5.2 Black Hole Quasi-Normal Modes

Merging binaries of compact objects produce black holes that are initially highly deformed. As mentioned before, the energy in the deformation is emitted as a superposition of exponentially damped sinusoidal gravitational waves called quasi-normal modes. There are infinitely many quasi-normal modes although only a small number of them (≈ 10) might be excited with nonnegligible amplitudes in the process of the merger of compact objects [26.87]. As a consequence of the *no hair theorem*, the complex frequencies of modes all depend only on the merged black hole's mass and spin angular momentum. (A black hole could, in principle, also have an electric charge. Astrophysical black holes, which are the ones of interest to us, are, however, expected not to have any residual charge.)

The modes are indexed by integers similar to the spherical harmonic indices (ℓ, m), $\ell = 2, 3, \ldots$, $m = -\ell, \ldots, \ell$, as also an overtone index n that takes values $n = 0, 1, 2, \ldots$. By definition, $n = 0$ is the least damped mode and called the *fundamental* mode. In most problems, it is sufficient to consider the fundamental mode. Gravitational waves emitted during quasi-normal mode ringing of a black hole are given by [26.88]

$$h_+ - \mathrm{i}h_\times = \sum_{\ell=2}^{\infty} \sum_{m=-\ell}^{\ell} Y_{-2}^{\ell m}(\iota, \phi) h_{\ell m} ,$$

$$h_{\ell m} = \frac{\alpha_{\ell m} M}{D} \mathrm{e}^{-\mathrm{i}\omega_{\ell m} t - t/\tau_{\ell m}} ,$$

where (ι, ϕ) refer to the colatitude and the azimuth angle at which the radiation is emitted from the black hole and $Y_{-2}^{\ell m}$ are -2 spin-weighted spherical harmonics. Although the mode frequencies $\omega_{\ell m}$ and damping times $\tau_{\ell m}$ are functions of only the final black hole's mass and spin, the amplitudes $\alpha_{\ell m}$ of the excited modes depend quite critically on the parameters of the progenitor binary and in particular on its mass ratio and spins. As a result, by detecting quasi-normal modes, even when the binary itself is not visible to a detector (because the inspiral phase lies outside the sensitive band), it might still be possible to measure the parameters of the progenitor binary [26.89, 90]. Moreover, accurate measurement of the various mode frequencies could provide ambiguous evidence for the existence of black holes as the mode frequencies of other compact objects (e.g., neutron stars) will depend on more than two parameters [26.91]. Consistency of the modes with one another will provide smoking gun evidence for black holes.

Quasi-normal modes could also be used to test general relativity by comparing the predictions of numerical solutions to Einstein's equations with the measured spectrum of the modes and their amplitudes [26.91, 92]. In particular, it should be possible to measure the mass of the system before and after merger and compute the *mass loss* to gravitational waves and see how that compares with predictions of general relativity [26.89], to *test the no-hair theorem* by measuring the complex frequencies of different modes and verifying if they depend on extra degrees of freedom other than the object's mass and spin and confirming that the merged object is actually a black hole and not a naked singularity [26.93].

26.5.3 Neutron Stars

Neutron stars in isolation but with a time-varying quadrupole moment can be good sources of gravitational waves. The birth of a neutron star in a supernova, neutron stars with mountains that rotate about an axis misaligned with the symmetry axis, accretion of matter onto a neutron star from a companion star, and trans-

fer of energy from a differentially rotating neutron-star core to crust, can all produce gravitational waves that are potentially detectable. Any isolated body is bound to have a limited supply of energy available for radiation and so either most of the energy might be emitted in a burst resulting in a strong source that would be easily discernible (e.g., supernova), or the energy might leak out slowly over millions of years, giving a long-lived continuous, but weak, source of radiation (e.g., continuous waves from a neutron star whose spin axis is different from its axis of symmetry).

Supernovae: The Birth of Neutron Stars. Neutron stars are born in the aftermath of the collapse of a massive star or when the core of a white dwarf becomes more massive than the Chandrasekhar limit of $1.4M_{\odot}$. Both axisymmetric and nonaxisymmetric collapse can produce gravitational waves. In fact, the first gravitational wave detectors were built to detect galactic supernova and they are still among the most important sources. Supernovae produce the Universe's dust and heavy elements; their cores are laboratories of complex physical phenomena requiring general relativity, nuclear physics, magneto-hydrodynamics, neutrino viscosity, and transport, turbulence, etc., to model them. Much of the physics of supernovae is poorly understood: how nonaxisymmetric is the collapse? How much energy is converted to gravitational waves and over what time scale? What causes shock revival in supernovae that form a neutron star? Gravitational wave observations could provide some of the clues for solving these questions [26.94].

We can make a rough estimate of the amplitude of the radiation if we have a knowledge of the total energy in gravitational waves and the frequency of the emitted waves. For a galactic supernova ($r \approx 10\,\mathrm{kpc}$), assuming $\Delta E = 10^{-8} M_{\odot}$, $\Delta t = 10\,\mathrm{ms}$, and $f = 300\,\mathrm{Hz}$, the amplitude from (26.13) is $h \approx 10^{-21}$. Such an event would produce a characteristic amplitude $h_c \approx h/\sqrt{f} \approx 6\times 10^{-23}\,\mathrm{Hz}^{-1/2}$. Amplitudes this large would be detectable in advanced detectors, especially if we know the epoch of the event and its sky position. Supernovae occur only once in about 30 or 100 years in a galaxy like the Milky Way and so the prospect of observing a galactic supernova is not so bright. The supernova rate could be of order 1 per few years within 2 Mpc (see, e.g., *Ando* et al. [26.95]). At that distance advanced detectors will not be sensitive to supernovae, but ET will be.

Triaxial Neutron Stars. In many cases, one can think of a neutron star as a triaxial rotating body. A neutron star rotating at a frequency of f_{rot} emits gravitational waves at $f = 2f_{\mathrm{rot}}$. Using the quadrupole formula (26.11) it is straightforward to compute the amplitude of the radiation from a triaxial body to be

$$h_+(t) = h_0 \frac{1+\cos^2\iota}{2}\cos 2\phi(t)\,,$$
$$h_\times(t) = h_0 \cos\iota \sin 2\phi(t)\,, \tag{26.24}$$

where ι is the angle between the star's spin axis and the line of sight, and h_0 and $\phi(t)$ are the signal's amplitude and phase

$$h_0 = \frac{4\pi^2 G}{c^4}\frac{\epsilon I_{zz} f^2}{r}\,,$$
$$\phi(t) = \phi_0 + 2\pi f t + \sum_{k=1}^{n} \frac{f_k}{(k+1)!} t^{k+1}\,. \tag{26.25}$$

Here I_{zz} is the star's moment of inertia with respect to the rotation axis, the ellipticity ϵ is defined in terms of the principal moments of inertia as $\epsilon = (I_{xx} - I_{yy})/I_{zz}$, and r is the distance to the star. The pulsar's frequency will not be constant due to the loss of rotational energy to gravitational waves and so the rotational phase $\phi(t)$ is not just linear in time. A Taylor expansion of the phase that includes the quadratic and, if necessary, higher-order corrections is used to describe the phase evolution. In this model, ϕ_0, and f_k, $k = 1,\ldots,n$ are, respectively, the phase and the spin-down parameters in the rest frame of the star at the fiducial time $t = 0$, n being the number of spin-down parameters included in the model (e.g., $-f_1$ is the rate at which the star spins down).

At a spin frequency of $f_{\mathrm{rot}} = 100\,\mathrm{Hz}$ (a gravitational wave frequency of $f = 200\,\mathrm{Hz}$), for a source at 10 kpc and ellipticity $\epsilon = 10^{-6}$, the amplitude of the radiation is $h_0 \simeq 4.2\times 10^{-27}$. To compute the characteristic strain amplitude h_c produced by such a signal we must assume a time interval over which the signal is integrated; taking this to be 1 year, we obtain $h_c = h_0\sqrt{1\,\mathrm{yr}} = 2.3\times 10^{-23}\,\mathrm{Hz}^{-1/2}$ (Fig. 26.7, dash-dotted lines). The amplitude increases as the square of the spin frequency, so for the Crab pulsar (B0531+21, $r \simeq 2\,\mathrm{kpc}$) with a spin frequency of 30 Hz, $h_c \approx 10^{-23}\,\mathrm{Hz}^{-1/2}$, for the same ellipticity. If Crab's ellipticity is ten times higher, then it will be well within the reach of advanced detectors (Fig. 26.7b). However, it is not clear if neutron stars occur with ellipticities as large as 10^{-5}. Models are mostly able to compute the maximum ellipticity of neutron stars by subjecting the crust to breaking strains. Ellipticities computed in the literature range from values of 10^{-4} (for exotic equations of state) [26.96]

to 10^{-7} for conventional crustal shear [26.97]. Large toroidal magnetic fields of order 10^{15} G could produce ellipticities of order 10^{-6} [26.98], and accretion along magnetic fields could produce similar or an order of magnitude larger deformations [26.99]. The large range in possible eccentricities shows that gravitational wave observations could have a potentially high impact and science return in this area.

The radiation emitted in this case is roughly a monotonic signal, but the radiation back reaction and energy lost to electromagnetic radiation and particles could cause the frequency to slowly drift in time. Using (26.13), we can compute the luminosity for the example we considered above to be roughly $\mathcal{L} \approx 40 L_\odot$. If we assume that the rotational energy $E = I\omega^2/2$ of the star powers the radiation, then the maximum timescale over which the energy is exhausted is roughly $\tau \approx E/\mathcal{L} \approx 5\times10^8$ yr. Newly born neutron stars emitting in the gravitational window will take 100s of millions of years to exhaust their source of energy and are essentially continuous wave (CW) sources.

The motion of the detector with respect to the source (due to Earth's rotation and orbital motion) causes a modulation in the signal's amplitude and frequency. Encoded in this modulation is the signal's position on the sky, and so it will be possible to resolve the source's location subject to the Rayleigh criterion, $\delta\theta = 2\pi\lambda/L$, where $\delta\theta$ is the angular resolution, λ is the wavelength of the radiation, and $L = 2$ AU is the diameter of Earth's orbit. At a frequency of 100 Hz, $\delta\theta \approx 2''$. Moreover, the amplitude of the radiation will help constrain the product of the star's ellipticity and moment of inertia, which is one of the main ingredients that goes into determining its equation of state.

Observing a representative sample of the galactic population of neutron stars could transform astrophysical studies of compact objects. A catalog of CW sources would help understand the galactic supernova rate, their demographics will lead to insights on evolutionary scenarios of compact objects, their amplitudes and distances can be used to constrain the equation of state. (Gravitational wave observations alone cannot determine the distance to CW sources. However, it should be possible to measure the distance to a subset of them from radio observations.) Constraints on the range of ellipticity could help understand crustal strengths and test models of the structure and composition of neutron stars.

Pulsar Glitches and Magnetar Flares. Radio pulsars have very stable spins and their periods (P) change very slowly over time. Their small spin-down rate ($\dot{P} \lesssim 10^{-12}\,\mathrm{s}^{-1}$) is occasionally marked by a sudden increase in angular frequency Ω, an event that is called a *glitch* [26.100]. To date more than 300 glitches have been observed in about 100 pulsars [26.101] (Jodrell Bank Observatory maintains a glitch catalog at [26.102]). Vela (B0833-45) is a nearby ($r \approx 300$ pc) pulsar in which 16 glitches have been observed since its discovery in 1969. The magnitude of a glitch is measured in terms of the fractional change in the angular velocity, which is found to be in the range $\Delta\Omega/\Omega \approx 10^{-5}$–$10^{-11}$. Some time after a glitch, the pulsar returns to its regular spin-down evolution. The origin of pulsar glitches is not a settled matter, although the most favored explanation is that it is due to the transfer of angular momentum from a differentially rotating core to the crust.

Glitches are not the only transient phenomena observed in neutron stars. Sources of giant x- and gamma ray flashes are believed to be highly magnetized ($B \approx 10^{15}$–10^{16} G) neutron stars called *magnetars*. The source of high energy radiation is believed to be the decay of the magnetic field associated with a stellar quake. Star quakes could, in general, excite normal mode oscillations of the ultra dense core. The energy in the modes is emitted as gravitational waves with a characteristic frequency and decay time, similar to quasi-normal modes of black holes. Unlike black holes, however, the complex mode frequencies depend on the mass and radius of the neutron star. (Although the modes would depend on the spin of the star, at a first approximation it can be neglected.) One can obtain an estimate of the amplitude of gravitational waves from a glitch by noting that for a star with angular frequency Ω the glitch energy is $\Delta E_{\text{glitch}} \simeq I\Omega\Delta\Omega$, where I is the star's moment of inertia and $\Delta\Omega$ is the change in angular velocity. For Vela, whose spin frequency is $f_{\text{rot}} \simeq 11$ Hz, the largest glitch has $\Delta\Omega/\Omega \approx 3\times10^{-6}$ [26.101]. The corresponding glitch energy is $\Delta E_{\text{glitch}} \approx 8\times10^{-12} M_\odot$. If a tenth of this energy is available to normal modes, which can then emit gravitational waves at the fundamental mode frequency of $f \approx 2$ kHz, then we can expect a strain amplitude of

$$h_0 \approx \frac{1}{\pi f r}\sqrt{\frac{5\,G}{c^3}\frac{\Delta E}{\Delta t}} \simeq 2\times10^{-21}\left(\frac{300\,\text{pc}}{r}\right) \times \left(\frac{\Delta E}{10^{-12} M_\odot}\right)^{1/2}\left(\frac{2\,\text{kHz}}{f}\right)^{1/2},$$

where we have taken $\Delta t = f^{-1} = 0.5$ ms. Such a signal would produce a characteristic amplitude of $h_c =$

$h_0/\sqrt{f} \simeq 4.5 \times 10^{-23}\,\text{Hz}^{-1/2}$. Third generation detectors like ET should be able to detect such amplitudes in coincidence with radio observations. Figure 26.7 shows plausible characteristic amplitudes produced by normal modes of energy $10^{-12} M_\odot$, for mode frequencies in the range of 1.5–4 kHz and neutron star distances in the range 1 kpc to 10 kpc.

The frequency and decay time of normal modes have different dependences on the mass and radius of the star. Thus, by measuring the complex mode frequency of, say, the fundamental mode one can infer both the mass and radius of the star. The size of a neutron star of a given mass depends critically on the supranuclear equation of state of matter, which is currently highly uncertain. Gravitational wave observations of glitches will help to directly measure the equation of state of matter under extreme conditions of density, pressure, and magnetic fields, one of the most important unsolved problems in astronomy and nuclear physics.

Low Mass X-ray Binaries. Gravity in the vicinity of a compact star is so large that particles falling into it can become accelerated close to the speed of light. Charged particles accelerated in this way are responsible for intense flashes of x-rays in binary systems in which one of the stars is a compact object that accretes matter from a low-mass star, or a main sequence star, or a giant. Low-mass x-ray binaries (LMXBs) are systems where the donor star is less massive than the compact object. LMXBs are thought to have spun up millisecond pulsars.

LMXBs emit bursts of x-ray flashes that last for about 10 s and repeat once every few hours or days. Millisecond oscillations in burst intensity are observed in many LMXBs (for a review of LMXBs, see [26.103]). X-ray bursts are believed to be caused by thermonuclear burning of infalling matter, while oscillations are suspected to be caused by the neutron star spin. About 100 galactic LMXBs are known to date as also many extragalactic ones. Inferred spin frequencies of neutron stars in LMXBs in all these cases seem to have an upper limit of about 700 Hz [26.103]. The centrifugal breakup of neutron star spins for most equations of state is far higher, about 1500 Hz. It has, therefore, been a puzzle as to why neutron star spin frequencies are stalled. One reason for this could be that some mechanism operating in the neutron star emits gravitational waves and the resulting loss in angular momentum explains why neutron stars cannot be spun up beyond a certain frequency. If the accretion induced torque on a neutron star is balanced by the emission of gravitational waves, then the amplitude of gravitational waves must be

$$h_0 = 5 \times 10^{-27} \left(\frac{F_\text{X}}{10^{-8}\,\text{erg}\,\text{cm}^{-2}\,\text{s}^{-1}} \right)^{1/2} \times \left(\frac{300\,\text{Hz}}{f} \right)^{1/2} , \tag{26.26}$$

where F_X is the bolometric x-ray flux of the source, from which the mass accretion rate is inferred and thereby helps in computing the amplitude of gravitational waves [26.104, 105]. The exact mechanism causing the emission of gravitational waves can account for this amplitude if it can support an effective ellipticity of $\epsilon \approx 10^{-8}$ (cf. (26.25)). This *ellipticity* could be produced by a time-varying, accretion-induced quadrupole moment [26.105], or by relativistic instabilities (e.g., r-modes) [26.106], or by large toroidal magnetic fields [26.107]. It is assumed that the gravitational wave luminosity is related to the x-ray luminosity and the expected characteristic amplitude of gravitational radiation is shown in Fig. 26.7 for the well-known LMXB Sco-X1 as well as for the known galactic population of LMXBs. Advanced detectors should detect Sco-X1 if the system is losing all of its accreted angular momentum to gravitational waves (for a comprehensive analysis of the detectability of LMXBs, see [26.108]). Such a detection could help understand what is the mechanism behind limiting spin frequencies in LMXBs and in turn provide deeper insights into models of LMXBs.

26.5.4 Stochastic Backgrounds

Another class of continuous gravitational waves is an ever present stochastic gravitational wave background. Such a background might have been produced by physical processes in the early Universe (just as cosmic microwave background was produced at the big bang) [26.109] or by random superposition of a population of point sources throughout the Universe [26.110]. Detecting such a background and measuring its spectral features could provide insight into the physical processes in the very early Universe accessible in no other way or provide a census of sources at cosmological distances.

The strength of a stochastic background is measured not by the amplitude of the radiation (as we cannot follow the amplitudes of individual waves) but by the power spectrum that it produces. An equivalent, but more popular, way of characterizing the strength of

a background is to specify the energy density in the radiation, as a function of frequency, relative to the closure density of the Universe

$$\Omega_{\rm GW}(f) \equiv \frac{1}{\rho_{\rm C}} \frac{{\rm d}\rho_{\rm GW}(f)}{{\rm d}\log f} ,$$

where $\rho_{\rm C} = 3H_0^2c^2/(8\pi G) \simeq 8.8\times10^{-10}\,{\rm J\,m^{-3}}$ is the closure density, $H_0 = 72\,{\rm km\,s^{-1}\,Mpc^{-1}}$ is the present value of the Hubble parameter. The dimensionless quantity $\Omega_{\rm GW}$ is related to the strain power spectrum $S_{\rm GW}(f)$ (which has dimensions of $\rm Hz^{-1}$) by [26.111]

$$S_{\rm GW}(f) = \frac{3H_0^2}{10\pi^2} \frac{\Omega_{\rm GW}(f)}{f^3} .$$

Since detector sensitivities and source strengths are often compared on a plot of strain amplitudes, it is useful to define $h_{\rm GW}(f) \equiv \sqrt{S_{\rm GW}(f)}$. $h_{\rm GW}$ has units of $\rm Hz^{-1/2}$ but it is still not the relevant quantity if we wish to understand the detectability of a given stochastic background.

Stochastic signals are detected by cross-correlating the data from a network of two or more detectors over a bandwidth Δf for a time duration T. For a pair of detectors the SNR of the background power spectrum grows as $\sqrt{T\Delta f}$, and so the amplitude grows as $(T\Delta f)^{1/4}$. Thus, the characteristic amplitude of a background at frequency f is $h_{\rm c}(f) = h_{\rm GW}\,(T\Delta f)^{1/4}$. (Some authors, e.g., *Sesana* et al. [26.56], choose to divide their characteristic amplitude by the factor $(T\Delta f)^{1/4}$. This is because their characteristic amplitude is a measure of the PTA sensitivity while our characteristic amplitude refers to the signal strength. Note also that these and many other PTA authors use a dimensionless characteristic amplitude: $h_{\rm c}^{\rm PTA} = \sqrt{f}h_{\rm c}^{\rm Here}$.) At 100 Hz, where ground-based detectors have their best sensitivity, the characteristic amplitude for an integration period of 1 yr and bandwidth $\Delta f = f$ is

$$h_{\rm c}(f) \simeq 3.0\times10^{-24} \left(\frac{T\Delta f}{3\times10^9}\right)^{1/4} \left(\frac{f}{100\,{\rm Hz}}\right)^{-3/2} \times \left(\frac{\Omega_{\rm GW}}{10^{-9}}\right)^{1/2} {\rm Hz}^{-1/2} .$$

Figure 26.7a plots (dotted lines) $h_{\rm c}(f)$ for several values of $\Omega_{\rm GW}$ assumed to be independent of f. Advanced ground-based detectors should detect $\Omega_{\rm GW} \geq 10^{-9}$.

In the case of PTA the detection technique is essentially similar. Instead of just a pair of detectors one looks at the timing residuals of many stable millisecond pulsars to improve the sensitivity. For an integration time of 5 yr and 20 ms pulsars, PTA could reach a sensitivity level of

$$h_{\rm c}(f) \simeq 2.4\times10^{-11} \left(\frac{T\Delta f}{1.6}\right)^{1/4} \left(\frac{f}{6\,{\rm nHz}}\right)^{-3/2} \times \left(\frac{\Omega_{\rm GW}}{2.5\times10^{-10}}\right)^{1/2} {\rm Hz}^{-1/2}; .$$

This corresponds to the dimensionless amplitude of $h_{\rm c}^{\rm PTA} \simeq 2\times10^{-15}$ at a frequency of $(5\,{\rm yr})^{-1} \simeq 6\times10^{-9}\,{\rm Hz}$.

Populations of Point Sources. The most certain source of stochastic gravitational wave background is the one produced by the galactic white dwarf binary population. White dwarf binaries with orbital periods in the range of few hours to few minutes are abundant in the galaxy. The combined effect of 100s of millions such systems is a stochastic background radiation. This white dwarf binary (WDB) background should be visible in the frequency range of 0.1–2 mHz in LISA (see Fig. 26.7) [26.112]. This would correspond to energy density in gravitational waves of $\Omega_{\rm GW} = 10^{-12}$ at 1 mHz [26.113]. Some close white dwarf binaries, AM CVn systems, and ultra compact x-ray binaries should be detectable above the confusion background of WDB as shown in [26.112].

Inspiralling compact binaries at cosmological distances will also cause a confusion background. Astrophysical observations guarantee the presence of two such populations: BNS, which should be observable in ground-based detectors, and binary supermassive black holes, which should be observable by PTAs. In both cases, the lack of precise knowledge about the underlying population of sources and their coalescence rate as a function of redshift makes it hard to predict the precise strength of the background. *Regimbau* and *Mandic* estimate that the BNS population could produce a background strength of [26.114, 115]

$$\Omega_{\rm GW}^{\rm BNS}(f) \simeq 2.5\times10^{-10} \left(\frac{f}{100\,{\rm Hz}}\right)^{2/3} .$$

This is out of reach of advanced detectors but ET should be able to detect such a background quite easily. It should be noted, however, that unlike WDB population in LISA, BNS population will not cause a confusion background above the detector noise spectral density [26.116]. *Sesana* [26.117] considers a number of different mechanisms for the formation and evolution of supermassive black hole binaries (SMBBH) and

computes a median energy density of

$$\Omega_{\mathrm{GW}}^{\mathrm{SMBBH}}(f) \simeq 2\times10^{-10}\left(\frac{f}{10\,\mathrm{nHz}}\right)^{2/3} .$$

This is only a factor of about 10 in Ω (which translates to a factor of 3 in amplitude) below the current best upper limits [26.118] and could be reachable within the next 5–10 years.

Stochastic background could be produced by any population of point sources including stellar mass BBH, supernovae, and magnetars. In most cases the uncertainty in event rates is so high that it is difficult to predict reliable estimates (for reviews, see [26.110, 114]).

Early Universe. Physical processes in the very early Universe can lead to a stochastic gravitational wave background. On generic grounds, we should expect that just as electromagnetic stochastic radiation (cosmic microwave background) was produced at the birth of the Universe, a background of gravitational radiation was also generated. While electromagnetic radiation was in thermal equilibrium with relativistic particles for about 300 000 years after the big bang, gravitational waves, due to their weak interaction with matter, decouple from all particles and radiation a tiny fraction of a second after the birth of the Universe. They should, therefore, carry the signature of physical processes when quantum gravity effects were important. Observing relic gravitons from the early Universe is undoubtedly the most important goal for gravitational astronomy.

The primary mechanism for the generation of primordial gravitational waves is the parametric amplification (by the background gravitational field) of gravitational waves generated by quantum fluctuations in the inflationary era [26.109, 119, 120]. This is the same mechanism that is believed to have produced the scalar density perturbations that led to the formation of large-scale structure in the Universe. The standard de Sitter inflationary model predicts that the energy density Ω_{GW} is independent of frequency for $f < f_0$, where $f_0 \simeq 10^{-16}$ Hz, raising as a power-law for $f < f_0$ [26.110]

$$\Omega_{\mathrm{GW}}^{\mathrm{Inflation}} = \begin{cases} \Omega_0\left(\dfrac{f}{f_0}\right)^{-2}, & f \le f_0 \\ \Omega_0, & f > f_0 , \end{cases} ,$$

where the value of Ω_0 is uncertain and could be in the range $10^{-13} \le \Omega_0 \le 10^{-14}$. The transition frequency f_0 corresponds to the horizon scale at the time of matter and radiation equality redshifted to the current epoch. Gravitational waves with $f > f_0$ today were much smaller than the horizon scale at matter radiation equality and hence were not amplified. COBE observations place a bound of $\Omega_{\mathrm{GW}} \simeq 2\times10^{-12}$ (for $H_0 = 72\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$). This is at a level that might be detectable by ET and LISA. The more popular slow roll inflation predicts background density two orders of magnitude smaller and hence unreachable by ground and space-based detectors or PTA.

Many other interesting sources of stochastic background have been studied, including cosmic strings, phase transitions in the early Universe and processes during the re-heating phase (see *Maggiore* [26.110] for a detailed account). Of these, cosmic strings are possibly the most interesting ones that could produce energy densities of order $\Omega_{\mathrm{GW}} \approx 10^{-8}$–$10^{-7}$ that is flat in the frequency range 10^{-8}–10^{10} Hz.

26.6 Conclusions

The next decade will witness the opening of the gravitational window for observational astronomy. Many sources of gravitational waves are multimessengers emitting intense x-rays, gamma rays, radio waves, optical radiation, and neutrinos. Observation of gravitational radiation from astronomical sources will undoubtedly help us learn a great deal about fundamental theories of nature, verify if astrophysical black holes have the properties predicted by general relativity, measure the large-scale geometrical and topological properties of the Universe, and infer the very early history of the Universe by detecting or constraining the stochastic background that might have been produced in the early Universe. As always, new windows of observation will undoubtedly reveal completely unexpected physical processes and astronomical phenomena, and this is perhaps where most of the exciting new discoveries will be made.

References

26.1 J.H. Taylor, L.A. Fowler, P.M. McCulloch: Measurements of general relativistic effects in the binary pulsar PSR1913+16, Nature **277**, 437–440 (1979)

26.2 J.M. Weisberg, J.H. Taylor: The relativistic binary pulsar b1913+16: Thirty years of observations and analysis, Binary Radio Pulsars. ASP Conf. Ser., Vol. 328, ed. by F.A. Rasio, I.H. Stairs (Astronomical Society of the Pacific, Aspen 2005) p. 25

26.3 P.S. Laplace: *Book X, Chapter VII, translated by N. Bowditch, original edition published in 1805*, A Treatise in Celestial Mechanics, Vol. IV (Chelsea, New York 1966)

26.4 L. Lorenz: On the identity of the vibrations of light with electrical currents, Philos. Mag. **34**, 287–301 (1867)

26.5 Wolfram Research, Long Hanborough, UK; http://scienceworld.wolfram.com/physics/LorenzGauge.html

26.6 J.B. Hartle: *An Introduction to Einstein's General Relativity* (Addison Wesley, San Francisco 2003)

26.7 D.J. Griffiths: *An Introduction of Electrodynamics*, 2nd edn. (Prentice Hall, Englewood Cliffs 19B9)

26.8 L.D. Landau, E.M. Lifshitz: *The Classical Theory of Fields* (Pergamon, Oxford 1971)

26.9 C.W. Misner, K.S. Thorne, J.A. Wheeler: *Gravitation* (Freeman, San Francisco 1973)

26.10 L. Blanchet: Gravitational radiation from post-Newtonian sources and inspiralling compact binaries, Liv. Rev. Relativ. **9**, 4 (2006)

26.11 Einstein A.: Näherungsweise Integration der Feldgleichungen der Gravitation, Sitzungsber. k. preuss. Akad. Wiss., 688–696 (1916)

26.12 R.A. Hulse, J.H. Taylor: Discovery of a pulsar in a binary system, Astrophys. J. **195**, L51–L53 (1975)

26.13 F.B. Estabrook, H.D. Wahlquist: Response of Doppler spacecraft tracking to gravitational radiation, Gen. Relativ. Gravit. **6**, 439–447 (1975)

26.14 J.W. Armstrong: Low-frequency gravitational wave searches using spacecraft doppler tracking, Liv. Rev. Relativ. **9**, 1 (2006)

26.15 B.S. Sathyaprakash, B.F. Schutz: Physics, Astrophysics and cosmology with gravitational waves, Liv. Rev. Rel. **12**, 2 (2009)

26.16 B.F. Schutz: Networks of gravitational wave detectors and three figures of merit, Class. Quantum Gravity **28**, 125023 (2011)

26.17 S. Fairhurst: Triangulation of gravitational wave sources with a network of detectors, New J. Phys. **13**(6), 069602 (2011)

26.18 J. Veitch, I. Mandel, B. Aylott, B. Farr, V. Raymond, C. Rodriguez, M. van der Sluys, V. Kalogera, A. Vecchio: Estimating parameters of coalescing compact binaries with proposed advanced detector networks, Phys. Rev. D **85**(10), 104045 (2012)

26.19 J. Weber: Detection and generation of gravitational waves, Phys. Rev. **117**, 306–313 (1960)

26.20 O.D. Aguiar: Past, present and future of the resonant-mass gravitational wave detectors, Res. Astron. Astrophys. **11**, 1–42 (2011)

26.21 L. Gottardi, A. de Waard, O. Usenko, G. Frossati, M. Podt, J. Flokstra, M. Bassan, V. Fafone, Y. Minenkov, A. Rocchi: Sensitivity of the spherical gravitational wave detector MiniGRAIL operating at 5K, Phys. Rev. D **76**(10), 102005 (2007)

26.22 O.D. Aguiar, J.J. Barroso, N.C. Carvalho, P.J. Castro, C.E. Cedeño, M. , C.F. da Silva Costa, J.C.N. de Araujo, E.F.D. Evangelista, S.R. Furtado, O.D. Miranda, P.H.R.S. Moraes, E.S. Pereira, P.R. Silveira, C. Stellati, N.F. Oliveira, Jr., X. Gratens, L.A.N. de Paula, S.T. de Souza, R.M. Marinho, Jr., F.G. Oliveira, C. Frajuca, F.S. Bortoli, R. Pires, D.F.A. Bessada, N.S. Magalhães, M.E.S. Alves, A.C. Fauth, R.P. Macedo, A. Saa, D.B. Tavares, C.S.S. Brandão, L.A. Andrade, G.F. Marranghello, C.B.M.H. Chirenti, G. Frossati, A. de Waard, M.E. Tobar, C.A. Costa, W.W. Johnson, J.A. de Freitas Pacheco, G.L. Pimentel: Status report of the Schenberg gravitational wave antenna, J. Phys. Conf. Ser. **363**(1), 012003 (2012)

26.23 M. Maggiore: *Gravitational Waves. Vol. 1: Theory and Experiments* (Oxford Univ. Press, Oxford 2007)

26.24 M.E. Gertsenshtein, V.I. Pustovoit: On the detection of low frequency gravitational waves, J. Exp. Theor. Phys. (USSR) **43**, 605–607 (1962)

26.25 S. Rowan, J. Hough: Laser interferometry for the detection of gravitational waves, J. Opt. A **7**, S257–S264 (2005)

26.26 B.F. Schutz, D. Nicholson, J.R. Shuttleworth, W.J. Watkins: Progress in gravitational wave detections, Proc. 6th Marcel Grossmann Meeting Recent Dev. Theor. Exp. Gen. Relativ., Gravit. Relativistic Field Theor., ed. by H. Sato, T. Nakamura (World Scientific, Singapore 1992) pp. 163–175

26.27 A. Abramovici, W.E. Althouse, R.W.P. Drever, Y. Gursel, S. Kawamura, F.J. Raab, D. Shoemaker, L. Sievers, R.E. Spero, K.S. Thorne, R.E. Vogt, R. Weiss, S.E. Whitcomb, M.E. Zucker: LIGO: The Laser interferometer gravitational wave observatory, Science **256**, 325–333 (1992)

26.28 B. Caron, et al.: The Virgo interferometer, Class. Quantum Gravity **14**, 1461–1469 (1997)

26.29 TAMA Collaboration: Current status of TAMA, Class. Quantum Gravity **19**, 1409–1419 (2002)

26.30 B. Willke, et al.: The GEO 600 gravitational wave detector, Class. Quantum Gravity **19**, 1377–1387 (2002)

26.31 B.P. Abbott, et al.: LIGO: The Laser Interferometer Gravitational-Wave Observatory, Rep. Prog. Phys. **72**(7), 076901 (2009)

26.32 B.P. Abbott, et al.: Searches for Gravitational Waves from known pulsars with science run 5 LIGO Data, Astrophys. J. **713**, 671–685 (2010)

26.33 B. Abbott, et al.: Implications for the Origin of GRB 070201 from LIGO Observations, Astrophys. J. **681**, 1419–1430 (2008)
26.34 B.P. Abbott, et al.: An upper limit on the stochastic gravitational-wave background of cosmological origin, Nature **460**, 990–994 (2009)
26.35 J. Aasi, et al.: Search for gravitational waves from binary black hole inspiral, merger, and ringdown in LIGO-Virgo data from 2009–2010, Phys. Rev. D **87**(2), 022002 (2013)
26.36 T. Bulik, K. Belczynski, A. Prestwich: IC10 X-1/NGC300 X-1: The very immediate progenitors of BH-BH binaries, Astrophys. J. **730**, 140 (2011)
26.37 J. Abadie, et al.: Search for gravitational waves associated with gamma-ray bursts during LIGO science run 6 and Virgo science runs 2 and 3, Astrophys. J. **760**, 12 (2012)
26.38 LIGO Scientific Collaboration, Virgo Collaboration: Prospects for localization of gravitational wave transients by the advanced LIGO and advanced virgo observatories (2013), arXiv e-prints
26.39 Sathyaprakash B., S. Fairhurst, B. Schutz, J. Veitch, S. Klimenko, D. Reitze, S. Whitcomb: Scientific benefits of moving one of LIGO Hanford detectors to India, Technical Report T1200219-v1 (LIGO Scientific Collaboration, Pasadena 2011)
26.40 S. Fairhurst: Source localization with an advanced gravitational wave detector network, Class. Quantum Gravity **28**(10), 105021 (2014)
26.41 LIGO Scientific Collaboration, California Institute of Technology, Pasadena USA; http://www.ligo.caltech.edu/~jzweizig/distribution/LSC_Data/
26.42 S. Fairhurst: Improved source localization with LIGO India, J. Phys. Conf. Ser. **484**, 012007 (2011)
26.43 M. Punturo, et al.: The Einstein Telescope: A third-generation gravitational wave observatory, Class. Quantum Gravity **27**, 194002 (2010)
26.44 The ET Science Team: *Einstein gravitational-wave telescope: Conceptual design study, Technical Report ET-0106A-10* (European Gravitational Observatory, 2011)
26.45 B. Sathyaprakash, et al.: Scientific objectives of Einstein Telescope, Class. Quantum Gravity **29**(12), 124013 (2012)
26.46 M. Pitkin, S. Reid, J. Hough: Gravitational wave detection by interferometry (ground and space), Liv. Rev. Relativ. **14**(5) (2011)
26.47 K. Danzmann: LISA – an ESA cornerstone mission for a gravitational wave observatory, Class. Quantum Gravity **14**, 1399 (1997)
26.48 N. Seto, S. Kawamura, T. Nakamura: Possibility of direct measurement of the acceleration of the universe using 0.1-Hz band laser interferometer gravitational wave antenna in space, Phys. Rev. Lett. **87**, 221103 (2001)
26.49 R.S. Foster, D.C. Backer: Constructing a pulsar timing array, Astrophys. J. **361**, 300–308 (1990)
26.50 G.H. Janssen, B.W. Stappers, M. Kramer, M. Purver, A. Jessner, I. Cognard: European pulsar timing array, 40 Years of Pulsars: Millisecond Pulsars, Magnetars and More. Am. Inst. Phys. Conf. Series, Vol. 983, ed. by C. Bassa, Z. Wang, A. Cumming, V.M. Kaspi (2008) pp. 633–635
26.51 R.N. Manchester: The Parkes pulsar timing array project, 40 Years of Pulsars: Millisecond Pulsars, Magnetars and More. Am. Inst. Phys. Conf. Series, Vol. 983, ed. by C. Bassa, Z. Wang, A. Cumming, V.M. Kaspi (2008) pp. 584–592
26.52 F. Jenet, L.S. Finn, J. Lazio, A. Lommen, M. McLaughlin, I. Stairs, D. Stinebring, J. Verbiest, A. Archibald, Z. Arzoumanian, D. Backer, J. Cordes, P. Demorest, R. Ferdman, P. Freire, M. Gonzalez, V. Kaspi, V. Kondratiev, D. Lorimer, R. Lynch, D. Nice, S. Ransom, R. Shannon, X. Siemens: The North American nanohertz observatory for gravitational waves (2009), ArXiv e-prints
26.53 G. Hobbs, A. Archibald, Z. Arzoumanian, D. Backer, M. Bailes, I. Cognard, M. Burgay, S. Burke-Spolaor, D. Champion, N.D.R. Bhat, M. Burgay, S. Burke-Spolaor, D. Champion, I. Cognard, W. Coles, J. Cordes, P. Demorest, G. Desvignes, R.D. Ferdman, L. Finn, P. Freire, M. Gonzalez, J. Hessels, A. Hotan, G. Janssen, F. Jenet, A. Jessner, C. Jordan, V. Kaspi, M. Kramer, V. Kondratiev, J. Lazio, K. Lazaridis, K.J. Lee, Y. Levin, A. Lommen, D. Lorimer, R. Lynch, A. Lyne, R. Manchester, M. McLaughlin, D. Nice, S. Oslowski, M. Pilia, A. Possenti, M. Purver, S. Ransom, J. Reynolds, S. Sanidas, J. Sarkissian, A. Sesana, R. Shannon, X. Siemens, I. Stairs, B. Stappers, D. Stinebring, G. Theureau, R. van Haasteren, W. van Straten, J.P.W. Verbiest, D.R.B. Yardley, X.P. You: The International Pulsar Timing Array project: using pulsars as a gravitational wave detector, Class. Quantum Gravity **27**(8), 084013 (2010)
26.54 F.A. Jenet, A. Lommen, S.L. Larson, L. Wen: Constraining the properties of supermassive black hole systems using pulsar timing: Application to 3C 66B, Astrophys. J. **606**, 799–803 (2004)
26.55 A. Sesana, A. Vecchio, M. Volonteri: Gravitational waves from resolvable massive black hole binary systems and observations with Pulsar Timing Arrays, Mon. Not. R. Astron. Soc. **394**, 2255–2265 (2009)
26.56 A. Sesana, A. Vecchio, C.N. Colacino: The stochastic gravitational-hack
wave background from massive black hole binary systems: implications for observations with pulsar timing arrays, Mon. Not. R. Astron. Soc. **390**, 192–209 (2008)
26.57 F.A. Jenet, G.B. Hobbs, W. van Straten, R.N. Manchester, M. Bailes, J.P.W. Verbiest, R.T. Edwards, A.W. Hotan, J.M. Sarkissian, S.M. Ord: Upper bounds on the low-frequency stochastic gravitational wave background from pulsar timing observations: Current limits and future prospects, Astrophys. J. **653**, 1571–1576 (2006)

26.58 B.G. Keating, A.G. Polnarev, N.J. Miller, D. Baskaran: The polarization of the cosmic microwave background due to primordial gravitational waves, Int. J. Mod. Phys. A **21**, 2459–2479 (2006)

26.59 The Planck Collaboration: Planck: The Scientific Programme, Technical Report ESA-SCI (2005) 1 (European Space Agency, Paris 2005)

26.60 M. Burgay, N. D'Amico, A. Possenti, R.N. Manchester, A.G. Lyne, B.C. Joshi, M.A. McLaughlin, M. Kramer, J.M. Sarkissian, F. Camilo, V. Kalogera, C. Kim, D.R. Lorimer: An increased estimate of the merger rate of double neutron stars from observations of a highly relativistic system, Nature **426**, 531–533 (2003)

26.61 A. Buonanno, T. Damour: Effective one-body approach to general relativistic two- body dynamics, Phys. Rev. D **59**, 084006 (1999)

26.62 A. Buonanno, T. Damour: Transition from inspiral to plunge in binary black hole coalescences, Phys. Rev. D **62**, 064015 (2000)

26.63 F. Pretorius: Evolution of binary black hole spacetimes, Phys. Rev. Lett. **95**, 121101 (2005)

26.64 B.S. Sathyaprakash, S.V. Dhurandhar: Choice of filters for the detection of gravitational waves from coalescing binaries, Phys. Rev. D **44**, 3819 (1991)

26.65 B.F. Schutz: Data processing, analysis and storage for interferometric antennas. In: *The Detection of Gravitational Waves*, ed. by D. Blair (Cambridge Univ. Press, Cambridge 1989) pp. 406–452

26.66 J. Abadie, et al.: Predictions for the rates of compact binary coalescences observable by ground-based gravitational-wave detectors, Class. Quantum Gravity **27**, 173001 (2010)

26.67 T. Bulik, K. Belczynski, A. Prestwich: IC10 X-1/NGC300 X-1: The very immediate progenitors of BH-BH binaries, Astrophys. J. **730**, 140 (2011)

26.68 K. Belczynski, M. Dominik, T. Bulik, R. O'Shaughnessy, C. Fryer, D.E. Holz: The effect of metallicity on the detection prospects for gravitational waves, astrophys. J. Lett. **715**, L138–L141 (2010)

26.69 R. Schödel, T. Ott, R. Genzel, R. Hofmann, M. Lehnert, A. Eckart, N. Mouawad, T. Alexander, M.J. Reid, R. Lenzen, M. Hartung, F. Lacombe, D. Rouan, E. Gendron, G. Rousset, A.-M. Lagrange, W. Brandner, N. Ageorges, C. Lidman, A.F.M. Moorwood, J. Spyromilio, N. Hubin, K.M. Menten: A Star in a 15.2 year orbit around the supermassive black hole at the center of the Milky Way, Nature **419**, 694–696 (2002)

26.70 A.M. Ghez, S. Salim, S.D. Hornstein, A. Tanner, J.R. Lu, M. Morris, E.E. Becklin, G. Duchene: Stellar orbits around the galactic center black hole, Astrophys. J. **620**, 744–757 (2005)

26.71 P. Amaro-Seoane, S. Aoudia, S. Babak, P. Binétruy, E. Berti, A. Bohé, C. Caprini, M. Colpi, N.J. Cornish, K. Danzmann, J.-F. Dufaux, J. Gair, O. Jennrich, P. Jetzer, A. Klein, R.N. Lang, A. Lobo, T. Littenberg, S.T. McWilliams, G. Nelemans, A. Petiteau, E.K. Porter, B.F. Schutz, A. Sesana, R. Stebbins, T. Sumner, M. Vallisneri, S. Vitale, M. Volonteri, H. Ward: eLISA: Astrophysics and cosmology in the millihertz regime (2012)

26.72 P. Amaro-Seoane, S. Aoudia, S. Babak, P. Binétruy, E. Berti, A. Bohé, C. Caprini, M. Colpi, N.J. Cornish, K. Danzmann, J.-F. Dufaux, J. Gair, O. Jennrich, P. Jetzer, A. Klein, R.N. Lang, A. Lobo, T. Littenberg, S.T. McWilliams, G. Nelemans, A. Petiteau, E.K. Porter, B.F. Schutz, A. Sesana, R. Stebbins, T. Sumner, M. Vallisneri, S. Vitale, M. Volonteri, H. Ward: Low-frequency gravitational-wave science with eLISA/NGO, Class. Quantum Gravity **29**, 124016 (2012)

26.73 K.G. Arun, S. Babak, E. Berti, N. Cornish, C. Cutler, J. Gair, S.A. Hughes, B.R. Iyer, R.N. Lang, I. Mandel, E.K. Porter, B.S. Sathyaprakash, S. Sinha, A.M. Sintes, M. Trias, C. Van Den Broeck, M. Volonteri: Massive black hole binary inspirals: Results from the LISA parameter estimation taskforce, Class. Quantum Gravity **26**, 094027 (2009)

26.74 S.A. Farrell, N.A. Webb, D. Barret, O. Godet, J.M. Rodrigues: An intermediate-mass black hole of over 500 solar masses in the galaxy ESO243-49, Nature **460**, 73–75 (2009)

26.75 A. Sesana, J. Gair, I. Mandel, A. Vecchio: Observing gravitational waves from the first generation of black holes, Astrophys. J. Lett. **698**, L129–L132 (2009)

26.76 J.R. Gair, I. Mandel, A. Sesana, A. Vecchio: Probing seed black holes using future gravitational-wave detectors, Class. Quantum Gravity **26**, 204009 (2009)

26.77 P. Amaro-Seoane, L. Santamaria: Detection of IMBHs with ground-based gravitational wave observatories: A biography of a binary of black holes, from birth to death, Astrophys. J. **722**, 1197–1206 (2010)

26.78 F.D. Ryan: Accuracy of estimating the multipole moments of a massive body from the gravitational waves of a binary inspiral, Phys. Rev. D **56**, 1845 (1997)

26.79 J.R. Gair, I. Mandel, A. Sesana, A. Vecchio: Probing seed black holes using future gravitational-wave detectors, Class. Quantum Gravity **26**(20), 204009 (2009)

26.80 B.F. Schutz: Determining the Hubble constant from gravitational wave observations, Nature **323**, 310 (1986)

26.81 N. Dalal, D.E. Holz, S.A. Hughes, B. Jain: Short GRB and binary black hole standard sirens as a probe of dark energy, Phys. Rev. D **74**, 063006 (2006)

26.82 B.S. Sathyaprakash, B.F. Schutz, C. Van Den Broeck: Cosmography with the Einstein telescope, Class. Quantum Gravity **27**, 215006 (2010)

26.83 S. Nissanke, D.E. Holz, S.A. Hughes, N. Dalal, J.L. Sievers: Exploring short gamma-ray bursts as

gravitational-wave standard sirens, Astrophys. J. **725**, 496–514 (2010)
26.84 H.-Y. Chen, D.E. Holz: GRB beaming and gravitational-wave observations (2012) ArXiv e-prints
26.85 E. Nakar: Short-hard gamma-ray bursts, Phys. Rep. **442**, 166–236 (2007)
26.86 C. Messenger, J. Read: Measuring a cosmological distance-redshift relationship using only gravitational wave observations of binary neutron star coalescences, Phys. Rev. Lett. **108**, 091101 (2012)
26.87 Y. Pan, A. Buonanno, M. Boyle, L.T. Buchman, L.E. Kidder, H.P. Pfeiffer, M.A. Scheel: Inspiralmerger-ringdown multipolar waveforms of nonspinning black-hole binaries using the effective-one-body formalism, Phys. Rev. D **84**, 124052 (2011)
26.88 E. Berti, V. Cardoso, A.O. Starinets: Quasinormal modes of black holes and black branes, Class. Quantum Gravity **26**, 163001 (2009)
26.89 I. Kamaretsos, M. Hannam, S. Husa, B.S. Sathyaprakash: Black-hole hair loss: Learning about binary progenitors from ringdown signals, Phys. Rev. D **85**, 024018 (2012)
26.90 I. Kamaretsos, M. Hannam, B. Sathyaprakash: Is black-hole ringdown a memory of its progenitor?, Phys. Rev. Lett. **109**, 141102 (2012)
26.91 O. Dreyer, B. Kelly, B. Krishnan, L.S. Finn, D. Garrison, R. Lopez-Aleman: Black hole spectroscopy: Testing general relativity through gravitational wave observations, Class. Quantum Gravity **21**, 787 (2004)
26.92 E. Berti, J. Cardoso, V. Cardoso, M. Cavagliá: Matched filtering and parameter estimation of ringdown waveforms, Phys. Rev. D **76**, 104044 (2007)
26.93 S. Gossan, J. Veitch, B.S. Sathyaprakash: Bayesian model selection for testing the no-hair theorem with black hole ringdowns, Phys. Rev. D **85**, 124056 (2012)
26.94 C.D. Ott: Probing the core-collapse supernova mechanism with gravitational waves, Class. Quantum Gravity **26**, 204015 (2009)
26.95 S. Ando, F. Beacom, H. Yüksel: Detection of Neutrinos from supernovae in nearby galaxies, Phys. Rev. Lett. **95**, 171101 (2005)
26.96 B.J. Owen: Maximum elastic deformations of compact stars with exotic equations of state, Phys. Rev. Lett. **95**, 211101 (2005)
26.97 G. Ushomirsky, C. Cutler, L. Bildsten: Deformations of accreting neutron star crusts and gravitational wave emission, Mon. Not. R. Astron. Soc. **319**, 902 (2000)
26.98 C. Cutler: Gravitational waves from neutron stars with large toroidal B fields, Phys. Rev. D **66**(8), 084025 (2002)
26.99 D.J.B. Payne, A. Melatos, E.S. Phinney: Gravitational waves from an accreting neutron star with a magnetic mountain, Astrophysics Gravitational Wave Sources. AIP Conf. Proc., Vol. 686, ed. by J.M. Centrella (American Institute of Physics, Melville 2003) pp. 92–95
26.100 N. Chamel, P. Haensel: Physics of Neutron Star Crusts, Liv. Rev. Relativ. **11**(10) (2008)
26.101 C.M. Espinoza, A.G. Lyne, B.W. Stappers, M. Kramer: A study of 315 glitches in the rotation of 102 pulsars, Mon. Not. R. Astron. Soc. **414**, 1679–1704 (2011)
26.102 Pulsar Glitches, Jodrell Bank Centre for Astrophysics, University of Manchester, UK; http://www.jb.man.ac.uk/pulsar/glitches.html
26.103 D. Chakrabarty: Millisecond pulsars in x-ray binaries, Binary Radio Pulsars. Astron. Soc. Pac. Conf. Series, Vol. 328, ed. by F.A. Rasio, I.H. Stairs (2005) p. 279
26.104 R.V. Wagoner: Gravitational radiation from accreting Neutron stars, Astrophys. J. **278**, 345–348 (1984)
26.105 L. Bildsten: Gravitational radiation and rotation of accreting neutron stars, Astrophys. J. Lett. **501**, L89 (1998)
26.106 N. Andersson, K.D. Kokkotas, N. Stergioulas: On the relevance of the R-mode instability for accreting neutron stars and white dwarfs, Astrophys. J. **516**, 307–314 (1999)
26.107 C. Cutler: Gravitational waves from neutron stars with large toroidal B fields, Phys. Rev. D **66**(8), 084025 (2002)
26.108 A.L. Watts, B. Krishnan, L. Bildsten, B.F. Schutz: Detecting gravitational wave emission from the known accreting neutron stars, Mon. Not. R. Astron. Soc. **389**, 839–868 (2008)
26.109 L.P. Grishchuk: Amplification of gravitational waves in an istropic universe, Sov. Phys. J. Exp. Theor. Phys. **40**, 409–415 (1975)
26.110 M. Maggiore: Gravitational wave experiments and early universe cosmology, Phys. Rep. **331**, 283 (2000)
26.111 B. Allen, J.D. Romano: Detecting a stochastic background of gravitational radiation: Signal processing strategies and sensitivities, Phys. Rev. D **59**, 102001 (1999)
26.112 G. Nelemans, L.R. Yungelson, S.F. Portegies Zwart: The gravitational wave signal from the galactic disk population of binaries containing two compact objects, Astron. Astrophys. **375**, 890 (2001)
26.113 E.S. Phinney: A practical theorem on gravitational wave backgrounds (2001)
26.114 T. Regimbau, V. Mandic: Astrophysical sources of stochastic gravitational-wave background, Class. Quantum Gravity **25**, 184018 (2008)
26.115 T. Regimbau: The astrophysical gravitational wave stochastic background, Res. Astron. Astrophys. **11**, 369–390 (2011)
26.116 T. Regimbau, T. Dent, W. Del Pozzo, S. Giampanis, T.G.F. Li, C. Robinson, C. Van Den Broeck, D. Meacher, C. Rodriguez, B.S. Sathyaprakash, K. Wójcik: A mock data challenge for the Ein-

stein gravitational-wave telescope, Phys. Rev. D **86**, 122001 (2012)
26.117 A. Sesana: Systematic investigation of the expected gravitational wave signal from supermassive black hole binaries in the pulsar timing band, Mon. Not. R. Astron. Soc. Lett. **433**(1), L1–L5 (2013)
26.118 R. van Haasteren, Y. Levin, G.H. Janssen, K. Lazaridis, M. Kramer, B.W. Stappers, G. Desvignes, M.B. Purver, A.G. Lyne, R.D. Ferdman, A. Jessner, I. Cognard, G. Theureau, N. D'Amico, A. Possenti, M. Burgay, A. Corongiu, J.W.T. Hessels, R. Smits, J.P.W. Verbiest: Placing limits on the stochastic gravitational-wave background using European Pulsar Timing Array data, Mon. Not. R. Astron. Soc. **414**, 3117–3128 (2011)
26.119 B. Allen: Stochastic gravity-wave background in inflationary-universe models, Phys. Rev. D **37**, 2078–2085 (1988)
26.120 L.P. Grishchuk, V.M. Lipunov, K.A. Postnov, M.E. Prokhorov, B.S. Sathyaprakash: Gravitational wave astronomy: In anticipation of first sources to be detected, Phys. Usp. **44**, 1 (2001)

27. Probing Dynamical Spacetimes with Gravitational Waves

Chris Van Den Broek

This decade will see the first direct detections of gravitational waves by observatories such as Advanced LIGO and Virgo. Among the prime sources are coalescences of binary neutron stars and black holes, which are ideal probes of dynamical spacetime. This will herald a new era in the empirical study of gravitation. For the first time, we will have access to the genuinely strong-field dynamics, where low-energy imprints of quantum gravity may well show up. In addition, we will be able to search for effects which might only make their presence known at large distance scales, such as the ones that gravitational waves must traverse in going from source to observer. Finally, coalescing binaries can be used as cosmic distance markers, to study the large-scale structure and evolution of the Universe.

With the advanced detector era fast approaching, concrete data analysis algorithms are being developed to look for deviations from general relativity in signals from coalescing binaries, taking into account the noisy detector output as well as the expectation that most sources will be near the threshold of detectability. Similarly, several practical methods have been proposed to use them for cosmology. We explain the state of the art, including the obstacles that still need to be overcome in order to make optimal use of the signals that will be detected. Although the emphasis will be on second-generation observatories, we will also discuss some of the science that could be done with future third-generation ground-based facilities such as Einstein Telescope, as well as with space-based detectors.

27.1 Overview

General relativity (GR) is a highly nonlinear, dynamical theory of gravitation. Yet, until the 1970s, almost all tests of GR were theoretically based on the behavior of test particles in a *static* gravitational field [27.1]. These include the perihelium precession of Mercury, deflection of star light by the Sun, and Shapiro time delay. The parameterized post-Newtonian (PPN) formalism (for an overview, see [27.2]) was developed to provide a systematic framework for these and other checks, by appropriately parameterizing various aspects of spacetime geometry viewed as a nonlinear superposition of contributions from, e.g., the Sun and the planets. Even

so, the most important early experiments did not require much more than an expansion of the Schwarzschild metric in $GM/(c^2r)$ up to the first few subleading terms, with M the mass and r the distance. Although excellent agreement with theory was obtained, the aspects of GR that were actually tested were somewhat limited, mostly amounting to the influence on the motion of test particles of low-order general-relativistic corrections to the Newtonian gravitational field.

The situation improved with the discovery of the Hulse–Taylor binary neutron star in 1974 [27.3]. One of the two stars can be observed electromagnetically as a pulsar, and from this signal it was inferred that the orbital motion of the binary changes as predicted by GR, assuming that orbital energy and angular momentum are being carried away by gravitational waves (GW). This was an event of historic significance, as it provided incontrovertible evidence of the dynamical nature of the gravitational field. Subsequently, similar and even more relativistic binary neutron stars were discovered, allowing for new tests of GR in a *post-Keplerian* framework [27.4]. Nevertheless, explaining the observed dissipative dynamics related to gravitational wave emission only requires a lowest-order approximation to GR in powers of v/c, with v a characteristic velocity.

Some of the most exciting aspects of general relativity still remain out of our empirical reach. What we would like to explore is the full nonlinear dynamics of the gravitational field itself, including its self-interaction. From this perspective, even the most relativistic binary neutron star system that is currently known, PSR J0737-3039 [27.4], is still in the relatively slowly varying, weak-field regime, with an orbital compactness of $GM/(c^2R) \simeq 4.4\times10^{-6}$ (where M is the total mass and R the orbital separation), and a typical orbital speed $v/c \simeq 2\times10^{-3}$. (For comparison, the surface gravity of the Sun is $\simeq 2\times10^{-6}$, and the orbital speed of the Earth as it moves around the Sun is $v/c \simeq 10^{-4}$.) By contrast, for a compact binary just prior to the final plunge and merger one has $GM/(c^2R) \approx 0.2$ and $v/c \approx 0.4$. This will bring us to the genuinely strong field, dynamical regime of general relativity, which in the foreseeable future will only be accessible by means of gravitational wave detectors.

The first detection of gravitational waves by the Advanced LIGO and Virgo detectors might happen as early as 2015, and certainly before the end of the decade [27.5]. Around 2020, a network of five large interferometric GW detectors will be in place: in addition to the two Advanced LIGO interferometers [27.6] and Advanced Virgo [27.7], there will be the Japanese KAGRA [27.8], and IndIGO in India [27.9]. There is also the smaller GEO-HF in Germany [27.10, 11]. These are usually referred to as second-generation detectors. A conceptual design study for a third-generation observatory called Einstein telescope (ET) was recently concluded, and there are plans for a space-based observatory called LISA [27.12]. There is a considerable body of literature on the projected capabilities of ET and LISA in probing the dynamics of gravity. Although attention will be given to these, our main focus in this chapter will be on what can be achieved with the upcoming advanced detectors. In particular, in the last few years, development has started of hands-on data analysis techniques for use on signals detected with second-generation observatories, properly taking into account the noisy detector output as well as the expectation that most sources will be near the threshold of detectability.

Since the advent of GR, a large number of alternative theories of gravity has been proposed; for a partial list, see [27.13]. Among these, GR tends to be the simplest and the most elegant; moreover, many of the alternatives are already strongly constrained by existing observations. Consequently, our primary aim will not be to seek confirmation of an alternative theory and measure its parameters; rather, we want to develop a test of GR itself. The testing should be as generic as we can make it, in the sense that if macroscopic deviations from GR exist, we want to find them even if they take a form that is yet to be envisaged, rather than looking inside a class of particular, existing alternative theories. That said, we will occasionaly mention the predictions of such alternative models to indicate the power of the probe that direct gravitational wave detection will provide us with.

Recently proposed tests of the strong-field dynamics broadly fall into two categories. One consists of checking for possible alternative polarization states beyond the two polarizations that GR predicts, and which might only make their presence known in the case of gravitational waves that were generated in the ultra-relativistic regime. The other category focuses on the coalescence process of compact binary objects (neutron stars and black holes) [27.2].

Searching for alternative polarizations started in earnest with the detailed studies made on the electromagnetically observed binary neutron stars. Here we will explain how one would go about looking for their signature in data from gravitational wave detectors, in particular using transient signals such as will be emitted by supernova explosions or, again, coalescing compact binaries. There have also been studies about how to use

stochastic gravitational waves for this purpose [27.14]; these could take the form of a primordial GW background, or they could be a *bath* of radiation caused by a large number of unresolvable astrophysical sources, such as the combined population of all compact binary coalescence events, or cosmic string cusps. Due to space limitations, here we will limit ourselves to resolvable transient sources. Although with a single interferometric detector one would not be able to tell the difference between, or even measure, additional polarizations, this does become possible with a *network* of detectors. In particular, it is possible to combine the outputs of multiple interferometers to construct a so-called *null stream*, which should be devoid of signals if the only polarization states present are the ones predicted by GR. More generally, one can have null streams which in addition to the usual tensor polarizations also exclude one or more of the alternative ones, allowing one to tell them apart. Here we will partially follow the recent discussions by *Chatziioannou* et al. [27.15], and by *Hayama* and *Nishizawa* [27.16].

The coalescence of compact binaries consists of three regimes: an adiabatic *inspiral*, the *merger* leading to the formation of a single black hole (or an exotic alternative object!), and the *ringdown* of the resulting object as it evolves toward a quiescent state. The inspiral regime is reasonably well understood thanks to the so-called post-Newtonian formalism [27.17], in which physical quantities such as energy and flux are expanded in powers of v/c. A test of GR could then take the form of identifying directly measurable quantities, such as post-Newtonian coefficients in an expansion of the orbital phase, which in GR are inter-related, and checking whether the predicted relationships really hold. This would constitute a very generic test of GR, in which no recourse needs to be taken to any particular alternative theory of gravity. Such a test was first proposed by *Arun* et al. [27.18–20]. Next, *Yunes* and *Pretorius* developed the *parameterized post-Einsteinian* (ppE) framework, which considerably generalized the family of waveforms used in [27.18–20] by allowing for a larger class of parameterized deformations [27.13, 21]. The basic idea of Arun et al. was implemented in a Bayesian way by *Li* et al. using waveforms in the ppE family [27.22, 23]. The latter approach focuses on hypothesis testing. This has the advantage that since for every detected sources the same *yes/no* question is asked, evidence for or against GR has a tendency to build up as information from an increasing number of detections is included.

Moving beyond the inspiral regime, the ringdown can be studied in a variety of ways. In particular, the *no hair theorem* can be tested, which says that in GR, the spacetime around a quiescent, electrically neutral black hole is determined uniquely by its mass and spin [27.24, 25]. The ringdown process can be modeled as perturbations on a fixed black hole spacetime, and the Einstein field equations impose relationships between the ringdown frequencies and damping times of the various modes that can get excited [27.26]. Verifying these interdependences amounts to testing the no hair theorem. Moreover, it has been shown that the *amplitudes* of the ringdown modes carry information about the masses and the spins of the binary compact object that merged to form a single black hole [27.27]. As shown earlier by *Ryan*, the no hair theorem can also be tested by monitoring the motion of a small object (a neutron star or a stellar mass black hole) around a very massive black hole or exotic object, effectively mapping out the latter's spacetime geometry [27.25].

As demonstrated by Schutz already in 1986, inspiraling and merging compact binaries can also be used as cosmic distances markers, or *standard sirens*, to probe the large-scale structure and evolution of the Universe [27.28]. A variety of techniques have been proposed to exploit this fact using the second-generation detectors, ET, and space-based detectors. The second-generation observatories will mainly give us information about the Hubble constant H_0; however, they will do so in a way that is completely independent of existing measurements, and in particular does not require the so-called cosmic distance ladder, with its potentially unknown systematic errors at every rung [27.29–31]. In the case of ET and space-based detectors, it is also possible to study the matter content of the Universe [27.32–36]. By now we know that the expansion of the Universe is accelerating [27.37, 38], which can be modeled heuristically by invoking a new substance called *dark energy*. An exciting prospect is probing the equation of state of dark energy with gravitational waves, again in a way that is independent of conventional observations.

This chapter is structured as follows. In Sect. 27.2 we discuss how one might look for alternative polarization states in transient GW signals, using a network of detectors. Next, we explain how the inspiral of compact binaries can be used to arrive at a very generic test of the strong-field dynamics of general relativity, including self-interaction (Sect. 27.3). The emphasis will be on second-generation detectors, where most sources will be near the threshold of detectability. As we shall

see, in the case of binary neutron stars, appropriate data analysis methods are already in place, which can be applied to raw data from the advanced detectors as soon as they become available. Binary black holes are dynamically far richer, but they also pose formidable data analysis problems. A discussion of merger and ringdown, and tests of the no hair theorem, will naturally bring us to third-generation ground-based observatories as well as space-based detectors (Sect. 27.4), and we will give an overview of what one might expect from them. In Sect. 27.5, we will briefly recall the basics of modern cosmology, and investigate what gravitational wave observations might have to say about the evolution of the Universe. A summary and conclusions are given in Sect. 27.6.

We will denote binary neutron stars by BNS, neutron star-black hole systems by NSBH, and binary black holes by BBH. The usual post-Newtonian notation will be employed, where *qPN order* with q integer or half-integer refers to $\mathcal{O}\left[(v/c)^{2q}\right]$ beyond leading order in expansions in v/c. Unless stated otherwise, we use units such that $G = c = 1$.

27.2 Alternative Polarization States

In the so-called transverse-traceless gauge, the metric perturbation only has spatial components, and for a signal propagating in the z direction in a coordinate system associated with unit vectors $(\hat{e}_x, \hat{e}_y, \hat{e}_z)$, it can be brought in the form [27.1]

$$h_{ij}^{\mathrm{TT}} = h_+ e_{ij}^+ + h_\times e_{ij}^\times \,, \tag{27.1}$$

with

$$e_{ij}^+ = \hat{e}_x \otimes \hat{e}_x - \hat{e}_y \otimes \hat{e}_y \,, \tag{27.2}$$

$$e_{ij}^\times = \hat{e}_x \otimes \hat{e}_y + \hat{e}_y \otimes \hat{e}_x \,. \tag{27.3}$$

h_+ and $h_\times$ are the magnitudes of the independent *plus* and *cross* tensor polarizations, respectively. The response to a gravitational wave of an L-shaped interferometric detector is a linear combination of these

$$h(t) = F_+ h_+(t) + F_\times h_\times(t) \,. \tag{27.4}$$

The *beam pattern functions* F_+, $F_\times$ depend on the sky position $\hat{\Omega} = (\theta, \phi)$ of the source

$$F_+ = \tfrac{1}{2}(1+\cos^2\theta)\cos 2\phi \,, \tag{27.5}$$

$$F_\times = -\cos\theta \sin 2\phi \,. \tag{27.6}$$

In metric theories of gravity, up to 6 degrees of freedom are allowed [27.2]; these are illustrated in Fig. 27.1. Other than the plus and cross polarizations, they include a scalar *breathing mode* ("b"), a scalar longitudinal mode ("ℓ"), and vectorial modes ("vx", "vy"). The full metric perturbation then takes the form (see [27.14] and references therein)

$$h_{ij} = h_+ e_{ij}^+ + h_\times e_{ij}^\times + h_\mathrm{b} e_{ij}^\mathrm{b} + h_\ell e_{ij}^\ell + h_{vx} e_{ij}^{vx} + h_{vy} e_{ij}^{vy} \,, \tag{27.7}$$

with

$$e^\mathrm{b} = \hat{e}_x \otimes \hat{e}_x + \hat{e}_y \otimes \hat{e}_y \,, \tag{27.8}$$

$$e^\ell = \sqrt{2}\hat{e}_z \otimes \hat{e}_z \,, \tag{27.9}$$

$$e^{vx} = \hat{e}_x \otimes \hat{e}_z + \hat{e}_z \otimes \hat{e}_x \,, \tag{27.10}$$

$$e^{vy} = \hat{e}_y \otimes \hat{e}_z + \hat{e}_z \otimes \hat{e}_y \,. \tag{27.11}$$

The full response of an interferometer to a signal containing all these polarization states is

$$h = F_+ h_+ + F_\times h_\times + F_\mathrm{b} h_\mathrm{b} + F_\ell h_\ell + F_{vx} h_{vx} + F_{vy} h_{vy} \,, \tag{27.12}$$

and

$$F_\mathrm{b} = -\tfrac{1}{2}\sin^2\theta\cos 2\phi \,, \tag{27.13}$$

$$F_\ell = \tfrac{1}{\sqrt{2}}\sin^2\theta\cos 2\phi, \tag{27.14}$$

$$F_{vx} = -\tfrac{1}{2}\sin 2\theta\cos 2\phi \,, \tag{27.15}$$

$$F_{vy} = \sin\theta\sin 2\phi \,. \tag{27.16}$$

Currently, observational constraints on additional polarization modes are limited. From the Hulse–Taylor double neutron star we know that the energy loss due to nontensor emission must be less than 1% [27.2, 16]. However, alternative polarizations might show up in more weak-field regimes and after having propagated over distances much larger than the characteristic scale of the Hulse–Taylor binary. Alternatively, they might only appear in situations where the source is far more relativistic, with high characteristic velocities v/c. In core collapse supernovae, radial velocities $v/c \approx 0.25$ are attained [27.39], which may excite longitudinal modes. In the case of binary inspiral, $v/c > 0.4$ is reached before the final plunge, which

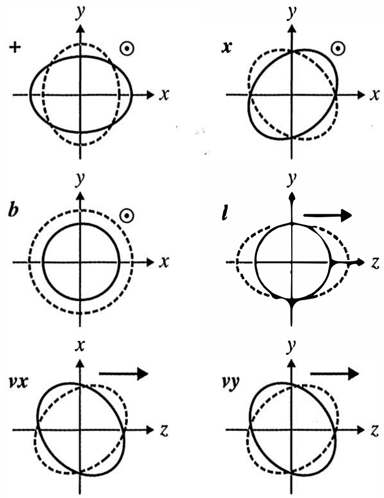


Fig. 27.1 In metric theories of gravity, up to six polarization states are allowed. At *top left* and *right*, we illustrate the transverse "+" and "×" tensor polarizations. At *middle left*, the transverse *breathing* mode is shown, and at *middle right* the longitudinal scalar mode. At *bottom left* and *right*, one has the vector modes (after [27.2])

(using Kepler's third law) corresponds to a gravitational wave frequency $f = c^3(v/c)^3/(\pi GM)$, with M the total mass. For binary neutron stars with component masses $(1.4, 1.4)\,M_\odot$ this approximately equals 1600 Hz, which is in the sensitivity band of ground-based detectors.

There are a number of alternative theories of gravity which predict nonstandard polarization states. To name but a few:

- Brans–Dicke theory is a scalar-tensor theory of gravity which has scalar modes [27.40, 41].
- Scalar modes also occur in Kaluza–Klein theory, where our 4D world arises after compactification of one or more extra spatial dimensions [27.42].
- Certain brane world models, such as the Dvali–Gabadadze–Porrati model in the self-accelerating branch, have all six modes above [27.43].

A *single* interferometric detector would not suffice to disentangle all these polarization states. To see this, consider a breathing mode (Fig. 27.1) impinging upon a detector, coming from a direction that is perpendicular to the plane of the interferometer. Then both detector arms would get lengthened and shortened in unison, but what an interferometer senses is the relative *difference* in arm length. Or, consider a breathing mode whose propagation direction corresponds to the orientation of one of the arms. Then this arm would be unaffected, while the other arm would still periodically get lenghtened and shortened, leading to a relative difference in arm lengths, which however would be indistinguishable from the effect of a gravitational wave with *plus* polarization. Hence, a *network* of interferometers is called for.

Consider D detectors at different positions on the Earth and whose noise is uncorrelated. A signal would reach the interferometers at different times. However, if one knew the sky position $\hat{\Omega}$, e.g., because of an electromagnetic counterpart to the gravitational wave signal as might be expected from a conveniently oriented BNS or NSBH event [27.44], then one would know how to time shift the outputs of the detectors to analyze the signal at a fixed time, say the arrival time at the Earth's center. Since from now on we assume $\hat{\Omega}$ to be known, we omit any explicit reference to it in expressions below. In each detector $A = 1, \ldots, D$, the output will take the form

$$\bar{d}^A(f) = \bar{h}^A(f) + \bar{n}^A(f)\,, \tag{27.17}$$

where each of the $\bar{h}_A$ takes the general form (27.12), and the $\bar{n}_A$ represent the noise in each of the detectors. The overbar indicates that (a) we will be considering the Fourier transforms of the relevant quantities, which are functions of frequency f rather than time t, and (b) the data streams have been divided by $\sqrt{S_A(f)}$, with $S_A(f)$ the *noise spectral density* (basically the variance of the noise as a function of frequency) for detector A. The latter ensures that we will not have to worry about differences in design and operation between the various detectors.

Equation (27.17) can be expressed in terms of the beam pattern functions

$$\bar{d}^A(f) = \bar{F}^A_a h_a(f) + \bar{n}^A(f)\,, \tag{27.18}$$

where $a = 1, \ldots, 6$ runs over the polarization states "+," "×," "b," "ℓ," "vx," and "vy," and summation over repeated indices is assumed. The first term in the right-hand side expresses the signal as a linear combination of five vectors in the D-dimensional space of detector outputs: $\bar{\boldsymbol{F}}_+$, $\bar{\boldsymbol{F}}_\times$, $\bar{\boldsymbol{F}}_{vx}$, $\bar{\boldsymbol{F}}_{vy}$, and $\bar{\boldsymbol{F}}_\ell$; indeed, from (27.13), (27.14), it is clear that $\bar{\boldsymbol{F}}_\ell = -\sqrt{2}\bar{\boldsymbol{F}}_{\rm b}$, so that one only has one independent vector pertaining to the scalar modes. Also note that the remaining vectors can be linearly independent only if $D \geq 5$.

In general relativity, only the tensor modes h_+ and $h_\times$ are present. Given three detectors (e.g., the two Ad-

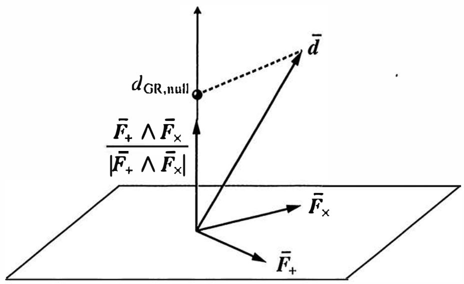


Fig. 27.2 An illustration of the construction of the null stream $d_{\mathrm{GR,null}}$ from a 3-detector output. The vector of outputs $\bar{d}$, and the beam pattern vectors $\bar{F}_+$ and $\bar{F}_\times$, live in a three-dimensional space. The null stream is obtained by projecting $\bar{d}$ onto the unit normal to the plane determined by $\bar{F}_+$, $\bar{F}_\times$. The projection is guaranteed not to contain tensorial polarization modes

vanced LIGOs and Advanced Virgo, which will be the first to take data), a *null stream* can be constructed by eliminating these modes from the output vector $\bar{d}$, following the original idea by *Gürsel* and *Tinto* [27.45]

$$d_{\mathrm{GR,\,null}} = \frac{\bar{F}_+ \wedge \bar{F}_\times}{|\bar{F}_+ \wedge \bar{F}_\times|} \cdot \bar{d}\,, \tag{27.19}$$

where $\bar{F}_+ \wedge \bar{F}_\times$ is the vector whose components in the space of detector outputs are

$$\epsilon^{ABC} \bar{F}_+^B \bar{F}_\times^C\,, \tag{27.20}$$

with ϵ^{ABC} the completely antisymmetric symbol, and here too summation over repeated indices is understood. It is not difficult to see that $d_{\mathrm{GR,\,null}}$ can only contain nonstandard polarizations; the tensor modes are projected out. This is illustrated in Fig. 27.2. Hence, if GR is correct, $d_{\mathrm{GR,\,null}}$ should not contain a signal. If, on the other hand, one or more of the alternative polarizations h_b, h_ℓ, h_{vx}, h_{vy} are present, then they will show up as a statistical excess in the null stream.

Sometime after 2017, the Japanese KAGRA will become active, and there will be four detectors, so that $D = 4$. This will allow for the construction of two null streams which in addition to the tensor modes will also be devoid of e.g., one of the two vector modes and one of the two scalar modes

$${}^{(4)}d^1_{\mathrm{GR,\,null}} = \frac{\bar{F}_+ \wedge \bar{F}_\times \wedge \bar{F}_{vx}}{|\bar{F}_+ \wedge \bar{F}_\times \wedge \bar{F}_{vx}|} \cdot \bar{d}\,, \tag{27.21}$$

$${}^{(4)}d^2_{\mathrm{GR,\,null}} = \frac{\bar{F}_+ \wedge \bar{F}_\times \wedge \bar{F}_\ell}{|\bar{F}_+ \wedge \bar{F}_\times \wedge \bar{F}_\ell|} \cdot \bar{d}\,, \tag{27.22}$$

where the wedge product is defined analogously to (27.20), but now using the four-dimensional antisymmetric symbol ϵ^{ABCD}. Note that for $D = 4$, there cannot be a third independent null stream which also excludes the tensor modes.

Finally, around the end of the decade, IndIGO may also be taking data, so that $D = 5$. In that case three null streams can be constructed that exclude the tensor modes

$${}^{(5)}d^1_{\mathrm{GR,\,null}} = \frac{\bar{F}_+ \wedge \bar{F}_\times \wedge \bar{F}_{vx} \wedge \bar{F}_{vy}}{|\bar{F}_+ \wedge \bar{F}_\times \wedge \bar{F}_{vx} \wedge \bar{F}_{vy}|} \cdot \bar{d}\,, \tag{27.23}$$

$${}^{(5)}d^2_{\mathrm{GR,\,null}} = \frac{\bar{F}_+ \wedge \bar{F}_\times \wedge \bar{F}_{vx} \wedge \bar{F}_\ell}{|\bar{F}_+ \wedge \bar{F}_\times \wedge \bar{F}_{vx} \wedge \bar{F}_\ell|} \cdot \bar{d}, \tag{27.24}$$

$${}^{(5)}d^3_{\mathrm{GR,\,null}} = \frac{\bar{F}_+ \wedge \bar{F}_\times \wedge \bar{F}_{vy} \wedge \bar{F}_\ell}{|\bar{F}_+ \wedge \bar{F}_\times \wedge \bar{F}_{vy} \wedge \bar{F}_\ell|} \cdot \bar{d}\,, \tag{27.25}$$

If a theory that has scalar modes happens to be the right one, then there will be a signal in ${}^{(5)}d^1_{\mathrm{GR,\,null}}$ above. If there are vector modes, then they will show up in ${}^{(5)}d^2_{\mathrm{GR,null}}$ and/or ${}^{(5)}d^3_{\mathrm{GR,null}}$.

Hayama and *Nishizawa* showed how to reconstruct the polarization modes in the case where the number of detectors is at least the number of modes, based on the null stream idea [27.16]. As an illustration, they reconstructed a simulated longitudinal mode in Brans–Dicke theory. Such a mode might be emitted by a supernova explosion, in which radial velocities $v/c \approx 0.25$ are reached [27.39].

If a statistical excess is seen in one or more null streams, then one would like to match-filter them with template waveforms that allow for one or more alternative polarization states to obtain information about their physical content. Such waveform models were developed in the context of the (extended) ppE framework by *Chatziioannou* et al. [27.15], and we refer the reader to that paper for details. The original ppE framework will be discussed below.

27.3 Probing Gravitational Self-Interaction

27.3.1 The Regime of Late Inspiral

Within GR, especially the inspiral part of the coalescence process has been modeled in great detail using the post-Newtonian (PN) formalism (see [27.17] and references therein), in which quantities such as the conserved energy and flux are found as expansions in v/c, where $v(t)$ is a characteristic speed. During inspiral, the GW signals will carry a detailed imprint of the orbital motion. Indeed, the main contribution has a phase that is simply $2\Phi(t)$, with $\Phi(t)$ the orbital phase. Thus, the angular motion of the binary is directly encoded in the waveform's phase, and assuming quasi-circular inspiral, the radial motion follows from the instantaneous angular frequency $\omega(t) = \dot{\Phi}(t)$ through the relativistic version of Kepler's third law. If there are deviations from GR, the different emission mechanism and/or differences in the orbital motion will be encoded in the phase of the signal waveform, allowing us to probe the strong-field dynamics of gravity.

In this section, we shall employ the usual post-Newtonian notation, in which *qPN order*, with q an integer or half-integer, refers to contributions at $(v/c)^{2q}$ beyond leading order.

Up to a reference phase, the orbital phase takes the form [27.46, 47]

$$\Phi(v) = \left(\frac{v}{c}\right)^{-5} \sum_{n=0}^{\infty} \left[\varphi_n + \varphi_n^{(l)} \ln\left(\frac{v}{c}\right)\right] \left(\frac{v}{c}\right)^n . \tag{27.26}$$

In general relativity, the coefficients φ_n and $\varphi_n^{(l)}$ depend on the component masses m_1, m_2 and spins $\boldsymbol{S}_1$, $\boldsymbol{S}_2$ in a very specific way; these dependences are currently known up to $n = 7$. The different PN terms in the phasing formula arise from nonlinear multipole interactions as the wave propagates from the source's *near zone*, where gravitational fields are strong, to the *far zone*, where detection takes place. Specifically, the physical content of some of the contributions is as follows:

- φ_3 and φ_5 encode the interaction of the total (Arnowitt–Deser–Misner, or ADM [27.1]) mass-energy of the source with the quadrupole moment. The physical picture is that the quadrupolar waves scatter off the Schwarzschild curvature generated by the source. These contributions are referred to as gravitational wave *tails*. One of the early proposals toward testing nonlinear aspects of general relativity using gravitational waves was due to *Blanchet* and *Sathyaprakash*, who first discussed the possibility of measuring these tail effects [27.48, 49].
- Spin–orbit interactions also first make their appearance in φ_3, and the lowest-order spin–spin interactions occur in φ_4 [27.50].
- φ_6 includes the cubic nonlinear interactions in the scattering of gravitational waves due to the ADM mass-energy of the system [27.48, 49].

Thus, observations of these PN contributions would allow for penetrating tests of the nonlinear structure of general relativity.

It is worth noting that with binary pulsars, one can only constrain the conservative sector of the orbital dynamics to 1 PN order, and the dissipative sector to leading order; see, e.g., the discussion in [27.51] and references therein. Hence, when it comes to $\Phi(t)$, these observations do not fully constrain the 1 PN contribution. More generally, terms in (27.26) with $n > 0$ are only accessible with direct gravitational wave detection.

27.3.2 The Parameterized Post-Einsteinian Formalism

By now there is a large body of literature on alternative theories to general relativity, which will induce changes in the functional dependences of the φ_n, $\varphi_n^{(l)}$ on component masses and spins, or even introduce new powers of v/c in the phase expression, (27.26). For instance:

- The effect of a nonstandard dispersion relation (e.g., due to a nonzero graviton mass) would accumulate over the large distances which the signal has to travel to reach the detector, and would be visible in φ_2. Solar system dynamics bound the graviton's Compton wavelength as $\lambda_g \gtrsim 10^{12}$ km. Second-generation detectors will improve on this by a factor of a few; Einstein Telescope will probe $\lambda_g \gtrsim 10^{14}$ km, and LISA $\lambda_g \gtrsim 10^{16}$ km [27.52–58].
- Scalar-tensor theories add a term $\varphi_{\mathrm{ST}}(v/c)^{-7}$ to (27.26), due to dipolar emission. In Brans–Dicke theory, one has a dimensionless parameter ω_{BD} which leads to standard GR in the limit $\omega_{\mathrm{BD}} \to \infty$. The Solar system bound from the Cassini spacecraft is $\omega_{\mathrm{BD}} \gtrsim 40\,000$; LISA will improve on this by up to an order of magnitude [27.53, 54, 59–61].
- A variable Newton constant adds a term $\varphi_{G(t)}(v/c)^{-13}$ [27.62], and extra dimensions can also have this effect [27.63].

- Quadratic curvature terms in the Lagrangian modify φ_4 [27.64, 65]. The same is true of dynamical Chern–Simons theory [27.66]. Here the second-generation detectors could place a bound on a dimensionful parameter of $\xi^{1/4} \lesssim \mathcal{O}(10\text{–}100)$ km, 6 to 7 orders of magnitude better than the solar system constraint ($\xi^{1/4} \lesssim \mathcal{O}(10^8)$ km), and in this case also considerably better than LISA ($\xi^{1/4} \lesssim \mathcal{O}(10^5\text{–}10^6)$ km)!

Quadratic curvature terms arise in string theory compactifications [27.67, 68], and dynamical Chern–Simons theory can be motivated both from string theory [27.69, 70] and loop quantum gravity [27.71, 72], and also arises in effective field theories of inflation [27.73]. Their effects on the phase at $(v/c)^4$ beyond leading order will only become visible when v/c is large. This is the regime we will be interested in here.

Yunes and *Pretorius* established the so-called ppE framework as a way both to classify alternative theories of gravity, and to provide template waveforms to search for violations of GR with gravitational wave detectors [27.13]. Their proposal involves both the phase and the amplitude of gravitational waves. However, since we are mostly concerned with second-generation detectors which for the expected stellar mass sources will not be very sensitive to changes in the amplitude [27.74, 75], we will focus on the phase. Instead of using the expression (27.26) for the inspiral phase, the authors of [27.13] proposed the following ansatz (again up to some reference phase)

$$\Phi(v) = \sum_{n=0}^{N} \left[\phi_n + \phi_n^{(l)} \ln\left(\frac{v}{c}\right)\right] \left(\frac{v}{c}\right)^{b_n} . \qquad (27.27)$$

Here, the b_n and ϕ_n, $\phi_n^{(l)}$ are meant to be completely free parameters. The above phase reduces to the one predicted by GR, (27.26), for $b_n = -5, -4, \ldots$ and when the phase coefficients have the standard dependences on component masses and spins: $\phi_n = \varphi_n(m_1, m_2, S_1, S_2)$, $\phi_n^{(l)} = \varphi_n^{(l)}(m_1, m_2, S_1, S_2)$. Yunes and Pretorius also showed how a variety of alternative theories of gravity in the literature can be obtained by making appropriate choices for the b_n and ϕ_n. Now, in the case of second-generation detectors, the form (27.27) may not be the most convenient one as far as data analysis is concerned. Indeed, even in the presence of a pure GR signal and using trial waveforms with the above phase, probability distributions arising from measurements of b_n and ϕ_n might peak at the correct values for very high SNRs, but probably not for signals at the threshold of detectability and in the presence of a considerable amount of noise, as is expected for most detections in second-generation observatories.

Yunes and others calculated the phase for a great variety of alternative theories, and in each case found the b_n to be integer; see the examples and references above. It then makes sense to write

$$\Phi(v) = \sum_{n=-2}^{N} \left[\phi_n + \phi_n^{(l)} \ln\left(\frac{v}{c}\right)\right] \left(\frac{v}{c}\right)^{n-5} , \qquad (27.28)$$

where we let the leading-order term be $(v/c)^{-7}$ to allow for dipolar emission. This time only the ϕ_n and $\phi_n^{(l)}$ are free parameters.

If there are too many free parameters to be determined, the measurement accuracy of *all* of the parameters will be adversely affected, and we would still like to reduce the number of free ϕ_n, $\phi_n^{(l)}$ in (27.28). Alternative theories that have a nonzero $n = -2$ contribution to the phase, such as scalar-tensor theories, can already be fairly well constrained using the electromagnetically observed binary pulsars [27.4]. With direct gravitational wave detection, the regime where we will be the most sensitive to GR violations is the one where v/c is large, which is out of reach for other observational methods. Hence we are mostly interested in new contributions to the phase with a power of v/c greater than or equal to -5. For this reason, below we will set $\phi_{-2} = \phi_{-1} = 0$.

27.3.3 A Generic Test of General Relativity with Inspiraling Compact Binaries: The TIGER Method

In probing the strong-field dynamics, one would like to be sensitive to almost *any* departure from general relativity, also through mechanisms that have yet to be envisaged. Hence what is needed is a test of GR that is as generic as possible. The possibility of such a test was first put forward by *Arun* et al. in [27.18–20], and the idea is illustrated in Fig. 27.3. If for simplicity we assume that the component objects have zero spins, then the GR values of the coefficients ϕ_n, $\phi_n^{(l)}$ in (27.28) only depend on the component masses (m_1, m_2). Hence only two of them are independent, and tests of GR could be performed by comparing any three of them and checking for consistency. (Needless to say, this does not mean that a *completely* generic test of GR is possible. In this picture, in principle there could be a GR violation which somehow still causes the error bands of any triplet of phasing coefficients to have a common region of overlap, but at the same, wrong component masses. See also the discussion in [27.76].)

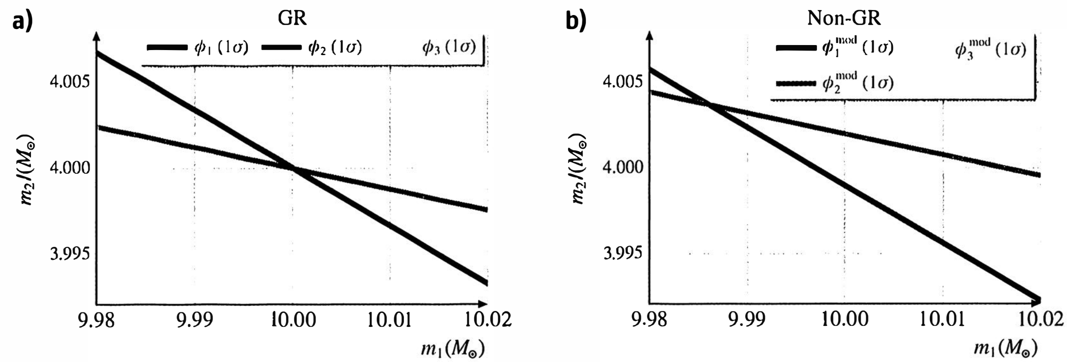


Fig.27.3a,b A schematic illustration of how one might set up a very generic test of GR (after [27.18–20]). The plots show the regions in the plane of the component masses (m_1, m_2) corresponding to the $1-\sigma$ measurement uncertainties on the coefficients (ϕ_1, ϕ_2, ϕ_3). **(a)** If GR is correct, there will be a common region of overlap at the true values of the masses (here 10 and 4 $M_\odot$). **(b)** If there is a deviation from GR and one or more of the ϕ_n do not have the dependences on masses that GR predicts, then there will be a mismatch

In practice, it is more convenient to make use of *Bayesian inference*. This involves the comparison of two hypotheses, namely the GR hypothesis $\mathcal{H}_{\mathrm{GR}}$, and $\mathcal{H}_{\mathrm{modGR}}$ which posits that GR is violated. In the present context, $\mathcal{H}_{\mathrm{GR}}$ will be the hypothesis that the ϕ_n, $\phi_n^{(l)}$ depend on both masses and spins in the standard way. Ideally, $\mathcal{H}_{\mathrm{modGR}}$ would be the negation of $\mathcal{H}_{\mathrm{modGR}}$, but this is impossible in principle to evaluate, as one cannot check the observed phase against *all* possible phase models that deviate from the GR family. Instead, we need to base our $\mathcal{H}_{\mathrm{modGR}}$ on a phase which allows for a finite-dimensional family of deviations.

Inspired by [27.18–20], we define $\mathcal{H}_{\mathrm{GR}}$ and $\mathcal{H}_{\mathrm{modGR}}$ as follows [27.22, 23]:

- $\mathcal{H}_{\mathrm{GR}}$ is the hypothesis that all the ϕ_n, $\phi_n^{(l)}$ have the functional dependence on component masses and spins as predicted by GR.
- $\mathcal{H}_{\mathrm{modGR}}$ is the hypothesis that *one or more* of the ϕ_n, $\phi_n^{(l)}$ (without specifying which) do not have this functional dependence, but all others do.

Given a detected inspiral signal in a stretch of data d, the question is now how these hypotheses are to be evaluated.

Suppose we would like to compare two hypotheses $\mathcal{H}_A$ and $\mathcal{H}_B$. First, on each of them we can apply Bayes' theorem [27.77]. For instance, for $\mathcal{H}_A$

$$P(\mathcal{H}_A|d,I) = \frac{P(d|\mathcal{H}_A,I)P(\mathcal{H}_A|I)}{P(d|I)} . \tag{27.29}$$

Here $P(\mathcal{H}_A|d,I)$ is the *posterior probability* of the hypothesis $\mathcal{H}_A$ given the data d and whatever additional information I we may hold, $P(\mathcal{H}_A|I)$ is the *prior probability* of the hypothesis, and $P(d|\mathcal{H}_A,I)$ is the *evidence* for $\mathcal{H}_A$, which can be written as

$$P(d|\mathcal{H}_A,I) = \int \mathrm{d}\boldsymbol{\theta}\, p(d|\mathcal{H}_A,\boldsymbol{\theta},I)p(\boldsymbol{\theta}|\mathcal{H}_A,I) . \tag{27.30}$$

In this expression, $p(\boldsymbol{\theta}|\mathcal{H}_A,I)$ is the prior probability density of the unknown parameter vector $\boldsymbol{\theta}$ within the model corresponding to $\mathcal{H}_A$, and $p(d|\mathcal{H}_A,\boldsymbol{\theta},I)$ is the likelihood function for the observation d, assuming the model $\mathcal{H}_A$ and given values of the parameters $\boldsymbol{\theta}$.

The function $p(d|\mathcal{H}_A,\boldsymbol{\theta},I)$ is what can be computed from the data. Let us assume that $\mathcal{H}_A$ corresponds to a particular gravitational wave signal model, $h_A(\boldsymbol{\theta};t)$. In the output of a gravitational wave detector $d(t)$, the signal will be combined with detector noise $n(t)$

$$\mathrm{d}(t) = n(t) + h_A(\boldsymbol{\theta};t) . \tag{27.31}$$

Let us assume that the noise is stationary and Gaussian; then its probability density distribution can be written as

$$p[n] = \mathcal{N}\mathrm{e}^{-(n|n)/2} , \tag{27.32}$$

where the square brackets in the left-hand side indicate that $p[n]$ is a *functional* of n, and $\mathcal{N}$ is a normalization factor. The inner product $(\cdot|\cdot)$ is defined as follows

$$(a|b) = 4\,\mathrm{Re}\int_0^\infty \frac{\tilde{a}^*(f)\tilde{b}(f)}{S_n(f)} , \tag{27.33}$$

with $\tilde{a}(f)$, $\tilde{b}(f)$ the Fourier transforms of functions $a(t)$, $b(t)$. The quantity $S_n(f)$ is called the *noise power spectral density*; comparing with (27.32), we see that it is essentially the variance of the noise as a function of frequency. Equations (27.31), (27.32), and (27.33) motivate the following form for the likelihood $p(d|\mathcal{H}_A, \boldsymbol{\theta}, I)$:

$$p(d|\mathcal{H}_A, \boldsymbol{\theta}, I) = \mathcal{N} \mathrm{e}^{-(d-h_A(\boldsymbol{\theta})|d-h_A(\boldsymbol{\theta}))/2} . \quad (27.34)$$

Indeed, when subtracting the signal from the detector output, the expectation is that only stationary, Gaussian noise remains.

Using (27.29) for both $\mathcal{H}_A$ and $\mathcal{H}_B$, one can construct an *odds ratio*

$$O_B^A \equiv \frac{P(\mathcal{H}_A|d, I)}{P(\mathcal{H}_B|d, I)} = \frac{P(\mathcal{H}_A|I)}{P(\mathcal{H}_B|I)} \frac{P(d|\mathcal{H}_A, I)}{P(d|\mathcal{H}_B, I)} , \quad (27.35)$$

where $P(\mathcal{H}_A|I)/P(\mathcal{H}_B|I)$ is the *prior odds* of the two hypotheses, i. e., the relative confidence we assign to the models before any observation has taken place. The ratio of evidences is called the *Bayes factor*, which can be computed from the data by using (27.30) and (27.34) for hypotheses $\mathcal{H}_A$ and $\mathcal{H}_B$

$$B_B^A \equiv \frac{P(d|\mathcal{H}_A, I)}{P(d|\mathcal{H}_B, I)} . \quad (27.36)$$

In the present context, the odds ratio of interest is

$$O_{\mathrm{GR}}^{\mathrm{modGR}} = \frac{P(\mathcal{H}_{\mathrm{modGR}}|d, I)}{P(\mathcal{H}_{\mathrm{GR}}|d, I)} = \frac{P(\mathcal{H}_{\mathrm{modGR}}|I)}{P(\mathcal{H}_{\mathrm{GR}}|I)} \frac{P(d|\mathcal{H}_{\mathrm{modGR}}, I)}{P(d|\mathcal{H}_{\mathrm{GR}}, I)} . \quad (27.37)$$

The evidence $P(d|\mathcal{H}_{\mathrm{GR}}, I)$ is computed by considering a large number of GR waveforms with different parameters $\boldsymbol{\theta}$ to map out the likelihood function $p(d|\mathcal{H}_{\mathrm{GR}}, \boldsymbol{\theta}, I)$, (27.34), which is then substituted into (27.30). However, the way $\mathcal{H}_{\mathrm{modGR}}$ is formulated, there is no waveform family associated with it, as there is no waveform model in which *one or more* of the ϕ_n, $\phi_n^{(l)}$ are different from their GR predictions.

To address this issue, we introduce the following *auxiliary hypotheses*:

> $H_{i_1 i_2 \cdots i_k}$ *is the hypothesis that the phasing coefficients* $\phi_{i_1}, \phi_{i_2}, \ldots, \phi_{i_k}$ *do not have the functional dependence on masses and spins as predicted by GR, but all other coefficients* ϕ_j, $j \notin \{i_1, i_2, \ldots, i_k\}$ *do have the dependence as in GR.*

Thus, for example, H_{12} is the hypothesis that ϕ_1 and ϕ_2 deviate from their GR values, but all other coefficients are as in GR. With each of the hypotheses above, we can associate a waveform model that can be used to test it. Let $\boldsymbol{\theta} = \{m_1, m_2, \boldsymbol{S}_1, \boldsymbol{S}_2, \ldots\}$ be the parameters occurring in the GR waveform, where m_1, m_2 are the component masses and $\boldsymbol{S}_1, \boldsymbol{S}_2$ the component spins; other parameters include the orientation of the orbital plane with respect to the line of sight, sky position, and distance. Then $H_{i_1 i_2 \cdots i_k}$ is tested by a waveform in which the independent parameters are

$$\{\boldsymbol{\theta}, \phi_{i_1}, \phi_{i_2}, \ldots, \phi_{i_k}\} , \quad (27.38)$$

i. e., the coefficients $\{\phi_{i_1}, \phi_{i_2}, \ldots, \phi_{i_k}\}$ are allowed to vary freely in addition to the other parameters.

The hypothesis we are really interested in is $\mathcal{H}_{\mathrm{modGR}}$ above, which posits that one or more of the ϕ_i differ from their GR values, without specifying which. But this corresponds to the logical *or* of the auxiliary hypotheses

$$\mathcal{H}_{\mathrm{modGR}} = \bigvee_{i_1 < i_2 < \cdots < i_k; k \le N_T} H_{i_1 i_2 \cdots i_k} . \quad (27.39)$$

Note that in practice, it will not be possible for computational reasons to consider all possible subsets of even the 10 known phasing coefficients; hence we limit ourselves to the subsets of $\{\phi_1, \phi_2, \ldots, \phi_{N_T}\}$, where $N_T \le 10$ is mainly set by computational resources. We will call the latter our *testing coefficients*.

To illustrate how the auxiliary hypotheses allow us to compute the odds ratio $\mathcal{O}_{\mathrm{GR}}^{\mathrm{modGR}}$ of (27.37), let us consider the case of just two testing coefficients, $\{\phi_1, \phi_2\}$. Then

$$\mathcal{H}_{\mathrm{modGR}} = H_1 \vee H_2 \vee H_{12} , \quad (27.40)$$

and the odds ratio becomes

$$O_{\mathrm{GR}}^{\mathrm{modGR}} = \frac{P(H_1 \vee H_2 \vee H_{12}|d, I)}{P(\mathcal{H}_{\mathrm{GR}}|d, I)} . \quad (27.41)$$

Now, the hypotheses H_1, H_2, and H_{12} are *logically disjoint*: the *and* of any two of them is false. Indeed, in H_1, ϕ_2 takes the GR value, but in H_2 it differs from it, as it does in H_{12}. Similarly, in H_2, ϕ_1 takes the GR value, but it differs from it in H_1 and in H_{12}. This implies

$$P(H_1 \vee H_2 \vee H_{12}|d, I) = P(H_1|d, I) + P(H_2|d, I) + P(H_{12}|d, I) \quad (27.42)$$

and hence

$$O_{\rm GR}^{\rm modGR} = \frac{P(H_1|d,I)}{P(\mathcal{H}_{\rm GR}|d,I)} + \frac{P(H_2|d,I)}{P(\mathcal{H}_{\rm GR}|d,I)} + \frac{P(H_{12}|d,I)}{P(\mathcal{H}_{\rm GR}|d,I)} . \quad (27.43)$$

Using Bayes' theorem (27.29) on each term, we get

$$O_{\rm GR}^{\rm modGR} = \frac{P(H_1|I)}{P(\mathcal{H}_{\rm GR}|I)} B_{\rm GR}^{1} + \frac{P(H_2|I)}{P(\mathcal{H}_{\rm GR}|I)} B_{\rm GR}^{2} + \frac{P(H_{12}|I)}{P(\mathcal{H}_{\rm GR}|I)} B_{\rm GR}^{12} , \quad (27.44)$$

where the Bayes factors $B_{\rm GR}^{1}$, $B_{\rm GR}^{2}$, and $B_{\rm GR}^{12}$ are given by

$$B_{\rm GR}^{1} = \frac{P(d|H_1,I)}{P(d|\mathcal{H}_{\rm GR},I)} , \quad B_{\rm GR}^{2} = \frac{P(d|H_2,I)}{P(d|\mathcal{H}_{\rm GR},I)} , \quad B_{\rm GR}^{12} = \frac{P(d|H_{12},I)}{P(d|\mathcal{H}_{\rm GR},I)} . \quad (27.45)$$

These can be computed from the data, as explained in the discussion leading up to (27.36). However, a choice will have to be made for the relative prior odds $P(H_1|I)/P(\mathcal{H}_{\rm GR}|I)$, $P(H_2|I)/P(\mathcal{H}_{\rm GR}|I)$, and $P(H_{12}|I)/P(\mathcal{H}_{\rm GR}|I)$. If one believed that the graviton has mass, then a deviation in ϕ_2 would be the thing to look for, and the auxiliary hypothesis H_2 should have more weight than either H_1 or H_{12}. On the other hand, one's favorite alternative theory might predict a violation in ϕ_1 instead, in which case H_1 should have more prior weight. Or, one might expect a GR violation to affect *all* phasing coefficients at the same time, so that H_{12} is *a priori* preferred. The method presented here is meant to find *generic* deviations from GR, with no preference for any particular alternative theory; consequently, we set

$$\frac{P(H_1|I)}{P(\mathcal{H}_{\rm GR}|I)} = \frac{P(H_2|I)}{P(\mathcal{H}_{\rm GR}|I)} = \frac{P(H_{12}|I)}{P(\mathcal{H}_{\rm GR}|I)} . \quad (27.46)$$

We will also need to specify the *overall* prior odds for $\mathcal{H}_{\rm modGR}$ against $\mathcal{H}_{\rm GR}$. Here we simply set

$$\frac{P(\mathcal{H}_{\rm modGR}|I)}{P(\mathcal{H}_{\rm GR}|I)} = \frac{P(H_1 \vee H_2 \vee H_{12}|I)}{P(\mathcal{H}_{\rm GR}|I)} = \alpha , \quad (27.47)$$

where the constant α will end up being an unimportant overall scaling factor in the odds ratio. Equations (27.46) and (27.47) imply

$$\frac{P(H_1|I)}{P(\mathcal{H}_{\rm GR}|I)} = \frac{P(H_2|I)}{P(\mathcal{H}_{\rm GR}|I)} = \frac{P(H_{12}|I)}{P(\mathcal{H}_{\rm GR}|I)} = \frac{\alpha}{3} , \quad (27.48)$$

and, except for the overall factor α, the final expression for the odds ratio reduces to a straightforward average of the Bayes factors for the auxiliary hypotheses against GR

$$O_{\rm GR}^{\rm modGR} = \frac{P(\mathcal{H}_{\rm modGR}|d,I)}{P(\mathcal{H}_{\rm GR}|d,I)} = \frac{\alpha}{3}\left(B_{\rm GR}^{1} + B_{\rm GR}^{2} + B_{\rm GR}^{12}\right) . \quad (27.49)$$

Thus, although there is no waveform model with which to directly test the hypothesis $\mathcal{H}_{\rm modGR}$, thanks to the auxiliary hypotheses it is nevertheless possible to compute its posterior probability relative to that of GR.

In the above example we used only two testing parameters, but in practice one will want to have more. With $N_{\rm T}$ testing parameters $\{\phi_1, \ldots, \phi_{N_{\rm T}}\}$ and making similar choices to (27.46) and (27.47), the odds ratio will again be proportional to an average of the Bayes factors for the auxiliary hypotheses against GR [27.22]

$$O_{\rm GR}^{\rm modGR} = \frac{\alpha}{2^{N_{\rm T}} - 1} \sum_{k=1}^{N_{\rm T}} \sum_{i_1 < i_2 < \cdots < i_k} B_{\rm GR}^{i_1 i_2 \cdots i_k} . \quad (27.50)$$

Combining data from multiple observed inspiral events will make for a far more robust test of GR compared to using just one detection. Suppose one has $\mathcal{N}$ independent detections in stretches of data $d_1, d_2, \ldots, d_{\mathcal{N}}$. Assuming these to be independent, it is not difficult to show that the odds ratio for the *catalog* of detections as a whole takes the form [27.22]

$$O_{\rm GR}^{\rm modGR} = \frac{P(\mathcal{H}_{\rm modGR}|d_1, d_2, \ldots, d_{\mathcal{N}}, I)}{P(\mathcal{H}_{\rm GR}|d_1, d_2, \ldots, d_{\mathcal{N}}, I)} = \frac{\alpha}{2^{N_{\rm T}} - 1} \sum_{k=1}^{N_{\rm T}} \sum_{i_1 < i_2 < \cdots < i_k} \prod_{A=1}^{\mathcal{N}} {}^{(A)}B_{\rm GR}^{i_1 i_2 \cdots i_k} , \quad (27.51)$$

i. e., for each auxiliary hypothesis, one multiplies together all the Bayes factors against GR for individual sources, ${}^{(A)}B_{\rm GR}^{i_1 i_2 \cdots i_k}$, after which one takes the average over all these hypotheses.

The algorithm described here was developed by *Li* et al. [27.22, 23]. It has been dubbed the TIGER method (*Test Infrastructure for GEneral Relativity*), and a hands-on data analysis pipeline for use on the

upcoming detections in Advanced LIGO and Virgo data has been developed based on this idea. It has a number of benefits:

- Unlike previous Bayesian treatments such as [27.21, 26], it addresses the question *Do one or more testing parameters deviate from their GR values?*, as opposed to *Do all of them deviate?*. Bayesian analysis naturally includes the idea of Occam's Razor in a quantitative way, and if the full non-GR model happens to have too many free parameters then one will be penalized for it [27.77].
- It is well suited to a situation where most sources are near the threshold of detectability. As shown in [27.22], if a GR violation is small, the Bayes factor for the *correct* auxiliary hypothesis (if any) will not always make the largest contribution to the odds ratio, as detector noise can obfuscate the precise nature of the GR violation. Even then, the GR hypothesis will typically be disfavored compared to one or more of the other auxiliary hypotheses, causing the GR violation to be detected after all.
- In combining information from multiple sources, it is not necessary that a GR violation manifests itself in the same way from one source to another. A deviation from GR could depend on mass, and on whatever additional charges might be present in the correct theory of gravity. However, in the above, the same *yes/no* question is asked for every source, and evidence for or against GR is allowed to build up as more and more sources get added.
- The method is not restricted to just the inspiral phase. It could equally well be applied to ringdown (as discussed below), or for that matter to alternative polarization states. All that is needed is a convenient parameterization of possible deviations from GR, such as provided by (generalizations of) the ppE formalism.

27.3.4 Accuracy in Probing the Strong-Field Dynamics with Second-Generation Detectors

Let us consider some examples to gauge how sensitive the TIGER method will be for particular (though heuristic) violations of GR, with the network of Advanced LIGO and Virgo detectors. In order to do this, one can produce simulated stationary, Gaussian detector noise, whose power spectral density (essentially the variance of the noise as a function of frequency) is in accordance with predictions for the Advanced LIGO and Virgo interferometers in their final configurations, projected for the 2019–2021 time frame [27.6, 7]. Simulated signals can be added to this simulated noise.

First we consider binary neutron stars. For such sources, the inspiral signal ends at high frequencies, and to good approximation one can assume that only this part of the coalescence process is visible in the frequency band where the detectors are sensitive. Moreover, in BNS systems the dimensionless intrinsic spins of the components are expected to be small: $cJ/(Gm^2) \lesssim 0.05$ [27.78], with J the spin and m the component mass. Finally, for most of the inspiral, neutron stars can be treated as point particles; finite size and matter effects will only be important at relatively high frequencies where the detectors are not very sensitive [27.79]. Thus, binary neutron star inspirals are relatively clean systems whose GW emission can be described by fairly simple waveform models [27.80]. Indeed, a hands-on data analysis pipeline which starts from raw detector data and computes $\mathcal{O}_{\rm GR}^{\rm modGR}$ has already been developed.

To have a fair assessment of how the method will perform with second-generation detectors, the simulated BNS sources will have to be distributed in an astrophysically realistic way. We will assume the component masses to be uniform in the interval $[1, 2]\,M_\odot$. The normal to the inspiral plane, and the sky position, are taken from a uniform distribution on the sphere, and sources are distributed uniformly in volume. Distances are between 100 and 400 Mpc; the former is the distance within which one can realistically expect one inspiral event every two years, and the latter is approximately the largest distance at which an optimally oriented and positioned system is still visible with Advanced LIGO [27.5]. Finally, many simulated catalogs of 15 sources each are produced.

We will be interested in GR violations that affect contributions to the phase (27.28) with $n > 0$; as mentioned earlier, we expect novel effects to show up for large v/c. Therefore, let us choose as testing coefficients the set $\{\phi_1, \phi_2, \phi_3\}$, leading to $2^3 - 1 = 7$ auxiliary hypotheses that need to be compared with $\mathcal{H}_{\rm GR}$ in order to compute the odds ratio $\mathcal{O}_{\rm GR}^{\rm modGR}$.

A priori, one would expect $\mathcal{O}_{\rm GR}^{\rm modGR} > 1$ if GR is violated, and $\mathcal{O}_{\rm GR}^{\rm modGR} < 1$ if it is correct. However, detector noise can mimick GR violations, so that occasionally one will have $\mathcal{O}_{\rm GR}^{\rm modGR} > 1$ even when GR is in fact correct. To deal with this, one can compute odds ratios for a large number of catalogs of simulated sources whose emission is in accordance with GR, but having different parameter values within the above ranges, and see how

the odds ratio ends up being distributed. For a given kind of GR violation, one can similarly construct an odds ratio distribution for catalogs of simulated sources. If the non-GR distribution has significant overlap with the GR distribution, then the particular GR violation considered would not necessarily be detected with great confidence. If, on the other hand, the two distributions are perfectly separated then the GR violation will be incontrovertibly detectable.

As we have seen, nonlinearities related to *tail* effects first show up at 1.5 PN order, i. e., in ϕ_3, so that this contribution is of particular interest. In order to gauge how sensitive the method might be to GR violations at this order, in [27.22] a constant relative shift in ϕ_3 was considered: $\phi_3 = (1 + \delta\chi_3)\phi_3^{\mathrm{GR}}$. This was compared with the GR case, and it was found that for $\delta\chi_3 = 0.1$, there is complete separation between the odds ratio distributions for the GR and non-GR catalogs, so that a violation of this kind and size would certainly be discovered. This is shown in the top panel of Fig. 27.4.

What if a deviation from GR does not manifest itself as simple shifts in the phase coefficients? Naively one might think that a more general phase model as in (27.27) would then be needed to uncover the GR violation. To show that this is not the case, *Li* et al. [27.23] considered a heuristic violation of the form

$$\Phi^{\mathrm{GR}}(\nu) \to \Phi^{\mathrm{GR}}(\nu) + \beta\left(\frac{\nu}{c}\right)^{-6+M/\mathrm{M}_\odot}, \qquad (27.52)$$

with $\Phi^{\mathrm{GR}}(\nu)$ the GR phase. The prefactor β was chosen to be of the same order as the ϕ_n predicted by GR (see [27.23] for details), and M is the total mass, so that the power of ν/c in the extra term varies from effectively 0.5 PN to effectively 1.5 PN within the BNS mass range considered. However, in order to compute $\mathcal{O}_{\mathrm{GR}}^{\mathrm{modGR}}$, the phase model used was still that of (27.28) with integer n, and the testing parameters were $\{\phi_1, \phi_2, \phi_3\}$, as before. As shown in the bottom panel of Fig. 27.4, also in this eventuality the GR hypothesis $\mathcal{H}_{\mathrm{GR}}$ will be disfavored compared with one or more of the $H_{i_1 i_2 \cdots i_k}$. The separation between GR and non-GR source catalogs is complete. The odds ratio $\mathcal{O}_{\mathrm{GR}}^{\mathrm{modGR}}$ indeed provides a Bayesian realization of the basic idea sketched in Fig. 27.3, inspired by *Arun* et al. [27.18–20].

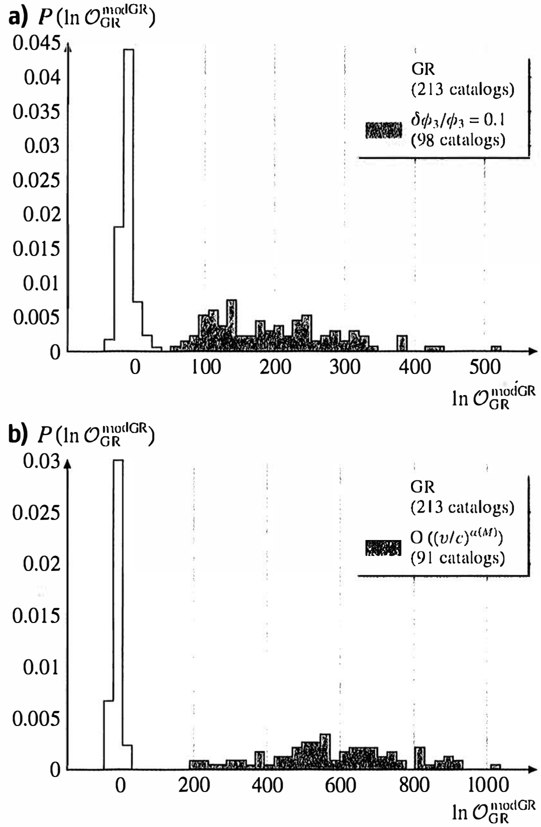


Fig. 27.4 (**a**) Distributions of log odds ratios for many catalogs of 15 BNS sources each, with and without a 10% shift in ϕ_3, the 1.5 PN phasing coefficient which contains the leading-order nonlinearities of GR related to tail effects (after [27.22]). The *light grey distribution* is for sources with pure GR emission; we see that mostly $\ln \mathcal{O}_{\mathrm{GR}}^{\mathrm{modGR}} < 0$, although noise will occasionally mimick a GR violation, causing the tail toward positive $\ln \mathcal{O}_{\mathrm{GR}}^{\mathrm{modGR}}$. The *grey distribution* is for sources with the 10% shift at 1.5 PN. The two distributions are perfectly separated, indicating that a violation of this type and magnitude will easily be detected. (**b**) The same for a violation which does not manifest itself as a simple shift in one or more of the phasing coefficients (see the main text for details), but $\mathcal{O}_{\mathrm{GR}}^{\mathrm{modGR}}$ is still computed in exactly the same way as before (after [27.23]). Here too, there is complete separation between GR and non-GR

27.3.5 Binary Neutron Stars Versus Binary Black Holes

As mentioned earlier, in the case of binary neutron stars, it is mostly only the inspiral part that is within the sensitivity band of the detector, so that we do not have to worry about the messy merger process. Finite size and matter effects mostly appear at high frequencies, where they have little impact. Neutron stars in binaries are expected to be relatively slowly spinning, and also

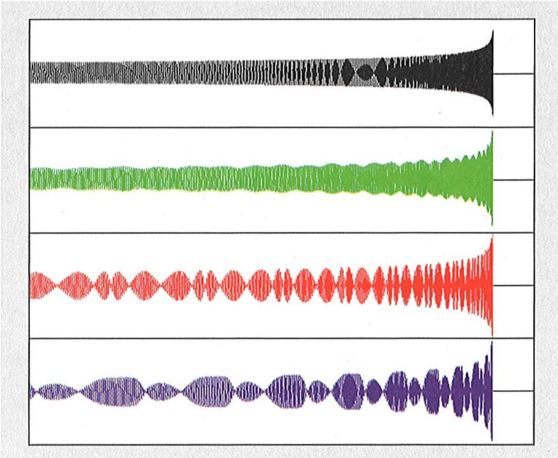


Fig. 27.5 For systems with no spins, the frequency and amplitude of gravitational waveforms increases in a steady *chirp* (*top*). If there are significant spins which are not aligned with each other and the orbital angular momentum, then one has precession of the orbital plane, causing modulation of the amplitude and frequency (courtesy of B.S. Sathyaprakash)

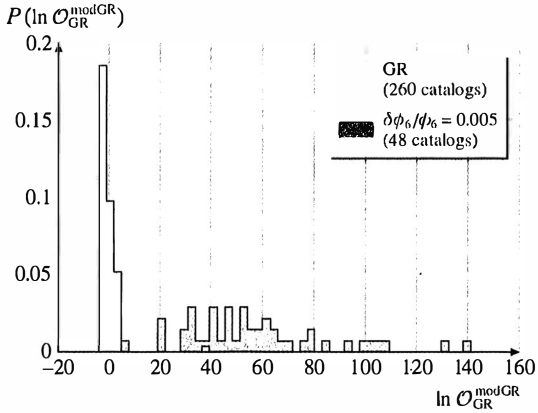


Fig. 27.6 An estimate by the authors of [27.22, 23] of how well one would be able to probe deviations from GR at high post-Newtonian order with binary black holes. As before, the *light grey histogram* is for catalogs of GR sources. The *grey histogram* is for sources with a 0.5% deviation at 3 PN order, where cubic nonlinear self-interaction of the gravitational field appears. The testing coefficients were $\{\phi_1, \phi_2, \phi_3, \phi_4\}$ and hence did not include the parameter ϕ_6 where the GR violation actually occurs; nevertheless, there is near-complete separation between GR and non-GR

this aspect can be dealt with. Consequently, a simple waveform model can be used for which data analysis algorithms are sufficiently fast, and a full data analysis pipeline for testing GR with BNS signals is already in place.

The situation is quite different for binary black holes. The frequency at which the inspiral terminates is roughly $c^3/(6^{3/2}\pi GM)$, with M the total mass. For a BBH with component masses of $(10, 10)\,M_\odot$, this is approximately 220 Hz, close to the frequency of ≈ 150 Hz where the detectors will be the most sensitive. Thus, the merger part of the signal, which is still not well modeled, will play a major role. Moreover, astrophysical black holes are expected to be fast-spinning, with dimensionless spins $cJ/(Gm^2) =$ 0.3–0.99 [27.81]. If the spins are not aligned with each other and the orbital angular momentum, then all three of these vectors will undergo precession during the inspiral phase [27.82, 83]. Since the unit vector $\hat{L}$ in the direction of orbital angular momentum is also the unit normal to the inspiral plane, the latter will undergo precession, and in extreme cases even a tumbling motion. This behavior is imprinted onto the gravitational wave emission through modulation of both the amplitude and the frequency of the waveform, as shown in Fig. 27.5. The rich dynamics that is unleashed in this way makes binary black holes far more interesting systems to study, but the more complicated signals also make the data analysis problem a great deal more difficult.

Large-scale numerical simulations provide us with accurate waveform models [27.84], but they take a long time to compute and cannot be used in data analysis, where many thousands of trial waveforms need to be compared with the data to arrive at accurate parameter estimation. On the other hand, semianalytic inspiral-merger-ringdown waveforms are under construction, which roughly fall into two categories. In the *Effective One-Body* formalism, the inspiral includes part of the final plunge, and a ringdown waveform can be *stitched* to it; the waveform as a whole can then be further *tuned* against numerical results [27.85, 86]. There is also a variety of phenomenological waveform models which are similarly improved using numerical predictions [27.87–91]. These are achieving matches $\gtrsim 0.99$ with numerically predicted signals; however, so far the only inspiral-merger-ringdown waveform with fully precessing spins is the one of [27.90, 91], which has been tuned against only a limited number of numerical waveforms.

Because of these difficulties, TIGER cannot yet be extended for use on BBH signals. However, the authors of [27.22, 23] made some rough estimates of what might be achievable once appropriate waveform

models are available. Figure 27.6 shows the log odds ratio distributions for catalogs of simulated BBH signals with GR emission, and with a 0.5% shift in ϕ_6, which encodes the cubic nonlinear interactions of the scattering of gravitational wave tails by the ADM mass-energy. In both cases, the inspiral-merger-ringdown waveform model used was the aligned-spin approximant of [27.89], and the set of testing parameters consisted of (the analogs of) only $\{\phi_1, \phi_2, \phi_3, \phi_4\}$. (Note that the number of auxiliary hypotheses $H_{i_1,\ldots,i_k}$ grows with the number of testing parameters N_T as $2^{N_T}-1$, and the data analysis problem can become computationally very costly if too many are used.) Hence the the coefficient containing the deviation was not among the testing parameters. However, even though both $\mathcal{H}_{GR}$ and the $H_{i_1,\ldots,i_k}$ are inconsistent with the signal, the waveform models of the latter have more freedom and can arrive at a closer fit, causing the GR hypothesis to be disfavored. And indeed, there is near-complete separation between GR and non-GR! On the other hand, precessing spins are bound to affect these results, in a way that is as yet unknown.

27.4 Testing the No Hair Theorem

In Newtonian theory, the gravitational potential Φ caused by a body with density ρ satisfies

$$\nabla^2\Phi = 4\pi G\rho \quad \text{in the interior}\,, \tag{27.53}$$

$$\nabla^2\Phi = 0 \quad \text{in the exterior}\,. \tag{27.54}$$

In the exterior, Φ can be expanded as

$$\Phi = -G\sum_{l,m}\frac{M_{lm}}{r^{l+1}}Y_{lm}(\theta,\phi)\,, \tag{27.55}$$

and the *multipole moments* M_{lm} are obtained by demanding consistency between the interior and exterior solutions. For axially symmetric objects, only terms with $m = 0$ contribute. The lowest-order multipole M_{00} is just the total mass of the body. By appropriately choosing the center of the coordinate system used, one can set $M_{10} = 0$. The next nontrivial multipole moment is M_{20}, the quadrupole moment, which has dimensions ML^2. The set of all multipole moments uniquely determines the shape of the potential Φ, and by measuring them one can study the properties of the mass distribution that gives rise to it.

In general relativity, the spacetimes outside bodies can similarly be described by a set of multipole moments. Spacetimes that are stationary, axisymmetric, reflection symmetric across the equatorial plane, and asymptotically flat – an example being the Kerr black hole – are characterized by *two* sets of multipole moments: mass multipole moments $M_0, M_2, M_4, \ldots$, and current (or spin) multipole moments $S_1, S_3, S_5, \ldots$. $M_0 = M$ is the mass, $S_1 = J$ is the spin, and M_2 is the mass quadrupole moment. Now, according to the no hair theorem [27.24], the multipole moments of quiescent black holes with the above properties satisfy

$$M_l + \mathrm{i}S_l = M(\mathrm{i}a)^l\,, \tag{27.56}$$

where $a = J/M$. Hence only two of them are independent: a quiescent black hole can be characterized completely by its mass M and spin J. Measuring any three of the multipole moments and checking consistency with the above relation would constitute a test of general relativity. The M_l and S_l have dimensions of $(\text{mass})^{l+1}$, and it is convenient to instead use dimensionless quantities $\mathfrak{m}_l = M_l/M^{l+1}$ and $\mathfrak{s}_l = S_l/M^{l+1}$, as we shall do below.

27.4.1 Ringdown

At the end of inspiral, binary neutron stars or black holes plunge toward each other to form a single, highly excited black hole, which will then undergo *ringdown* as it evolves to a quiescent, Kerr black hole. This process can be modeled as perturbations around a Kerr background, subject to the Einstein equations. For a black hole with mass M at a distance D, the *plus* and *cross* polarizations then take the form of damped sinusoids, the quasi-normal modes (QNMs) [27.92]

$$h_+(t) = \frac{M}{D}\sum_{l,m}A_{lm}Y_+^{lm}\mathrm{e}^{-t/\tau_{lm}}\cos(\omega_{lm}t - m\phi)\,, \tag{27.57}$$

$$h_\times(t) = -\frac{M}{D}\sum_{l,m}A_{lm}Y_\times^{lm}\mathrm{e}^{-t/\tau_{lm}}\sin(\omega_{lm}t - m\phi)\,, \tag{27.58}$$

with ϕ a phase offset, and Y_+^{lm}, $Y_\times^{lm}$ are linear combinations of spin-weighted spherical harmonics $_{-2}Y^{lm}$

[27.93],

$$Y_{+}^{lm}(\iota) = {}_{-2}Y^{lm}(\iota,0) + (-1)^l{}_{-2}Y^{l-m}(\iota,0)\,, \quad (27.59)$$

$$Y_{\times}^{lm}(\iota) = {}_{-2}Y^{lm}(\iota,0) - (-1)^l{}_{-2}Y^{l-m}(\iota,0)\,, \quad (27.60)$$

with ι the angle between the direction of the black hole's intrinsic angular momentum and the line of sight to the observer.

The damping times and mode frequencies $\tau_{lm}(M,J)$, $\omega_{lm}(M,J)$ in (27.57) and (27.58) only depend on the black hole mass M and its spin J [27.94, 95]. Hence, in general relativity only two of the τ_{lm} and ω_{lm} are independent, which opens up the possibility of a test of GR, similar to the one described above for the case of inspiral. This would effectively be a test of the no hair theorem. Indeed, the reason why frequencies and damping times only depend on these two quantities is that (a) the background spacetime around which one considers perturbations is assumed to be Kerr, and (b) the perturbative Einstein equations are assumed valid on this spacetime background, forcing relationships between damping frequencies and times.

Gossan et al. [27.26] demonstrated how one can exploit the interdependences of the τ_{lm}, ω_{lm} to test GR with Einstein Telescope as well as space-based detectors. For simplicity, they assumed that the spins of the progenitor objects were zero, in which case the amplitudes A_{lm} in (27.57) and (27.58) only depend on the symmetric mass ratio $\eta = m_1m_2/(m_1+m_2)^2$. Using data from numerical simulations in [27.96], they arrived at an analytic fit for the amplitudes A_{21}, A_{22}, A_{33}, and A_{44} of the four most dominant modes as a function of η, and the damping times $\tau_{lm}^{\mathrm{GR}}(M,J)$ and QNM frequencies $\omega_{lm}^{\mathrm{GR}}(M,J)$ predicted by GR were modeled using simple analytic fits from *Berti* et al. [27.97]. It was then assumed that the true damping times τ_{lm} and ω_{lm} might deviate from the GR prediction by dimensionless relative shifts $\Delta\hat{\tau}_{lm}$ and $\Delta\hat{\omega}_{lm}$, respectively

$$\tau_{lm} = (1+\Delta\hat{\tau}_{lm})\tau_{lm}^{\mathrm{GR}}(M,J)\,, \quad (27.61)$$

$$\omega_{lm} = (1+\Delta\hat{\omega}_{lm})\omega_{lm}^{\mathrm{GR}}(M,J)\,, \quad (27.62)$$

where in the case of GR, $\Delta\hat{\tau}_{lm} = \Delta\hat{\omega}_{lm} = 0$ for all l and m. The full parameter space for the *deviating* model $\mathcal{H}_{\mathrm{dev}}$ was then

$$\{\Delta\hat{\tau}_{lm}, \Delta\hat{\omega}_{lm}, \boldsymbol{\theta}\}\,, \quad (27.63)$$

with $\boldsymbol{\theta} = \{M, J, \ldots\}$ the parameters of the GR waveform. In practice, only a limited number of frequencies and damping times were allowed to be nonzero, leading to a parameter space

$$\{\Delta\hat{\omega}_{22}, \Delta\hat{\tau}_{22}, \Delta\hat{\omega}_{33}, \boldsymbol{\theta}\}\,. \quad (27.64)$$

With the second-generation detectors, it is unlikely that much more than the 22 mode will be distinguishable. As shown by *Kamaretsos* et al. [27.96], the situation is quite different for Einstein Telescope or a space-based detector such as LISA, where the 21, 22, 33, and 44 modes can all contribute signicantly to the signal-to-noise ratio. Numerical experiments were performed in which simulated signals were added to stationary, Gaussian noise following the projected noise power spectral densities of Einstein Telescope and LISA. The sensitivity to GR violations of the type (27.61) and (27.62) was then checked by two methods:

- *Direct parameter estimation.* Given data d and the signal model $\mathcal{H}_{\mathrm{dev}}$, the posterior probability distribution for the parameters $\boldsymbol{\lambda}$ of (27.63) is given by

$$p(\boldsymbol{\lambda}|d, \mathcal{H}_{\mathrm{dev}}, I) = \frac{p(d|\mathcal{H}_{\mathrm{dev}}, \boldsymbol{\lambda}, I)p(\boldsymbol{\lambda}|\mathcal{H}_{\mathrm{dev}}, I)}{p(d|\mathcal{H}_{\mathrm{dev}}, I)}\,, \quad (27.65)$$

where we have used Bayes' theorem. One has

$$p(d|\mathcal{H}_{\mathrm{dev}}, I) = \int \mathrm{d}\boldsymbol{\lambda}\, p(d|\mathcal{H}_{\mathrm{dev}}, \boldsymbol{\lambda}, I)p(\boldsymbol{\lambda}|\mathcal{H}_{\mathrm{dev}}, I) \quad (27.66)$$

with $p(\boldsymbol{\lambda}|\mathcal{H}_{\mathrm{dev}}, I)$ the prior distribution of the parameters. $p(d|\mathcal{H}_{\mathrm{dev}}, \boldsymbol{\lambda}, I)$ is the likelihood of the data given parameters $\boldsymbol{\lambda}$, as in (27.34)

$$p(d|\mathcal{H}_{\mathrm{dev}}, \boldsymbol{\lambda}, I) = \mathcal{N}\mathrm{e}^{-(d-h(\boldsymbol{\lambda})|d-h(\boldsymbol{\lambda}))/2}\,, \quad (27.67)$$

where $h(\boldsymbol{\lambda}; t)$ is the waveform family corresponding to $\mathcal{H}_{\mathrm{dev}}$. Posterior distributions for parameters like $\Delta\hat{\tau}_{lm}$ and $\Delta\hat{\omega}_{lm}$ are obtained by integrating the posterior probability density (27.65) over all the other parameters.
- *Model selection.* Here two models were considered: the GR model $\mathcal{H}_{\mathrm{GR}}$ in which $\Delta\hat{\tau}_{lm} = \Delta\hat{\omega}_{lm} = 0$ and only the $\boldsymbol{\theta}$ are free parameters, and the *deviating* model $\mathcal{H}_{\mathrm{dev}}$ in which the $\Delta\hat{\tau}_{lm}$ and $\Delta\hat{\omega}_{lm}$ are allowed to vary on top of the $\boldsymbol{\theta}$. One then computes

an odds ratio

$$O_{\rm GR}^{\rm dev} = \frac{P(\mathcal{H}_{\rm dev}|d,I)}{P(\mathcal{H}_{\rm GR}|d,I)} = \frac{P(\mathcal{H}_{\rm dev}|I)}{P(\mathcal{H}_{\rm GR}|I)}\frac{P(d|\mathcal{H}_{\rm dev},I)}{P(d|\mathcal{H}_{\rm GR},I)} . \tag{27.68}$$

The ratio of prior probabilities, $P(\mathcal{H}_{\rm dev}|I)/P(\mathcal{H}_{\rm GR}|I)$, is just a constant overall prefactor, which for convenience can be set to one.

With parameter estimation, it was found that a 10% deviation in ω_{22} would be clearly visible for a $500\,M_\odot$ black hole at 1.25 Gpc in ET, and for $10^6\,M_\odot$ and $10^8\,M_\odot$ at 1.25 Gpc and 10 Gpc, respectively, in LISA. $500M_\odot$ coalescences are expected to be rare within distances of 1.25 Gpc. By contrast, the quoted mass and distance range for LISA is expected to correspond to a detection rate of tens per year [27.98].

Bayesian model selection leads to rather better results. A 10% deviation in ω_{22} would be visible for $500M_\odot$ at $D_{\rm L} \simeq 6$ Gpc in Einstein Telescope, and for $10^6\,M_\odot$ at a similar distance with LISA. A 0.6 % deviation could be picked up at a redshift of 5 with a $10^8\,M_\odot$ source in LISA.

The above results are for black holes resulting from a binary with nonspinning components. However, in [27.27], it was shown that in the case where both components have spins, the ringdown mode amplitudes A_{lm} retain a memory not only of the progenitor's mass ratio, but also of spins.

We end this subsection with an important comment. In the odds ratio (27.68), *the hypothesis $\mathcal{H}_{dev}$ is not the equivalent of our $\mathcal{H}_{modGR}$ in the case of inspiral* (Sect. 27.3.3). To see this, denote the *testing parameters* by

$$(\phi_1, \phi_2, \phi_3) \equiv (\Delta\hat{\omega}_{22}, \Delta\hat{\tau}_{22}, \Delta\hat{\omega}_{33}) , \tag{27.69}$$

where in this case, $\phi_i = 0$ for $i = 1, 2, 3$ corresponds to GR. In the language of TIGER, one then has $\mathcal{H}_{\rm dev} = H_{123}$, the hypothesis in which all three parameters *at the same time* are different from the GR prediction. Indeed, in the *deviating* waveform model, all of the testing parameters are allowed to vary freely, but e.g., each of the hypersurfaces $\phi_i = 0$ have zero measure in the model's parameter space, and zero prior probability. In this notation and with the prior odds for $\mathcal{H}_{\rm dev}$ against $\mathcal{H}_{\rm GR}$ set to some arbitrary α

$$O_{\rm GR}^{\rm dev} = \alpha B_{\rm GR}^{123} . \tag{27.70}$$

The resolvability of anomalies in the testing parameters would no doubt improve if instead one were to compute an odds ratio

$$O_{\rm GR}^{\rm modGR} = \frac{P(\mathcal{H}_{\rm modGR}|d,I)}{P(\mathcal{H}_{\rm GR}|d,I)} \tag{27.71}$$

$$= \frac{\alpha}{2^3-1}\sum_{k=1}^{3}\sum_{i_1<\dots<i_k} B_{\rm GR}^{i_1\dots i_k} , \tag{27.72}$$

completely analogously to (27.50), possibly with a larger number of testing parameters than just $\{\Delta\hat{\omega}_{22}, \Delta\hat{\tau}_{22}, \Delta\hat{\omega}_{33}\}$. It would also be of great interest to see what happens if information from multiple sources is combined,

$$O_{\rm GR}^{\rm modGR} = \frac{P(\mathcal{H}_{\rm modGR}|d_1,\dots,d_{\mathcal{N}},I)}{P(\mathcal{H}_{\rm GR}|d_1,\dots,d_{\mathcal{N}},I)} , \tag{27.73}$$

analogously to (27.51).

27.4.2 Extreme Mass Ratio Inspirals

Extreme mass ratio inspirals (EMRIs) consist of a very massive black hole (or boson star [27.99], or naked singularity, ...) surrounded by a smaller object, which could be a neutron star or a stellar mass black hole. In the case of Einstein Telescope, target systems would have a massive component of a few hundred solar masses [27.100], while for space-based detectors the mass would be in the range $10^5 - 10^9\,M_\odot$ [27.101]. EMRIs provide another avenue to testing the no hair theorem: the orbits are expected to be extremely complicated, and in the case of LISA, there will be a large number of gravitational wave cycles within the detector's frequency band. As a consequence, the gravitational wave emission of the smaller object will bear a detailed imprint of the spacetime in the vicinity of the massive object.

Ryan was the first to evaluate the measurability of multipole moments using EMRI signals in Advanced LIGO and LISA [27.25]. With a number of simplifying assumptions – the most important one being circularity of the orbit of the smaller object, which moreover is taken to move in the equatorial plane – one can write down an expression for the phase in the Fourier domain explicitly showing the dependence on the multipoles $\mathfrak{m}_l$, $\mathfrak{s}_l$. In particular, $\mathfrak{s}_1$ first appears at 1.5 PN order, and $\mathfrak{m}_2$ at 2 PN. In the case of Advanced LIGO, Ryan's conclusion was that even assuming $m_1 = 30M_\odot$ (a very heavy stellar mass black hole) and $m_2 = 0.2$ (an im-

plausibly light neutron star), it would be hard to independently measure $\mathfrak{s}_l$ and $\mathfrak{m}_2$ with a single inspiral event near SNR threshold. For LISA, the situation is quite different. Assuming $m_1 = 10^5\,M_\odot$, $m_2 = 10\,M_\odot$, and SNR $= 100$, one obtains 1-σ measurement accuracies of $\Delta\mathfrak{s}_l \approx 10^{-4}$ and $\Delta\mathfrak{m}_2 \approx 1.5\times10^{-3}$. This would allow for a precision test of the no hair theorem.

Subsequent to Ryan's seminal paper, a number of authors have relaxed his assumptions. *Collins* and *Hughes* developed a treatment of multipole moments that is more appropriate than Ryan's in the strong-field regime [27.102] through the notion of *bumpy black holes*. The motion of the smaller object will not be quasi-circular; *Glampedakis* and *Babak* employed so-called kludge waveforms which encode the essentials of the orbital motion [27.103]. *Barack* and *Cutler* [27.104] showed that with kludge waveforms, results for the measurability of low-order multipole moments are qualitatively in keeping with those of Ryan. *Vigeland* and *Hughes* studied orbits around *bumpy black holes*, showing how a spacetime's bumps are imprinted onto the orbital frequencies [27.105]. Recently, *Vigeland* et al. studied bumpy black holes in alternative theories of gravity [27.106].

Most recently, *Rodriguez* et al. performed a more in-depth study of the possibility of using second-generation detectors to test the no hair theorem, also assuming a more reasonable mass for the lighter object ($1.4\,M_\odot$), and masses between $10\,M_\odot$ and $150\,M_\odot$ for the heavier one [27.107]. This work still mostly considered parameter estimation; it would be of great interest to cast the problem in terms of hypothesis testing, in which case results from multiple sources could be combined.

27.5 Probing the Large-Scale Structure of Spacetime

27.5.1 Binary Inspirals as Standard Sirens

Assuming that at large scales the Universe is homogeneous and isotropic, its line element can be put in the Friedmann–Lemaître–Robertson–Walker (FLRW) form [27.1]

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + a^2(t) \times\left[\frac{\mathrm{d}r^2}{1-kr^2} + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)\right], \tag{27.74}$$

where the entire dynamical content resides in the evolution of the scale factor $a(t)$. The constant k can be positive, zero, or negative, in which case the $t = const$ spatial hypersurfaces are hyperspherical, flat, or hyperboloidal, respectively. Given a homogeneous mass distribution ρ with pressure P, the Einstein equations reduce to two equations for $a(t)$, $\rho(t)$, and $P(t)$, called the Friedmann equations,

$$\left(\frac{\dot a}{a}\right)^2 = \frac{8\pi}{3}\rho - \frac{k}{a^2}, \tag{27.75}$$

$$\frac{\ddot a}{a} = -\frac{4\pi}{3}(\rho + 3P), \tag{27.76}$$

which can be combined to arrive at an equation for the time evolution of the density

$$\dot\rho = -3(\rho + P)\frac{\dot a}{a}. \tag{27.77}$$

This can be solved given an equation of state $P = P(\rho)$. In the case of pressureless dust (which can serve as a model for a sprinkling of galaxies), $P = 0$, and $\rho \propto a^{-3}$. For radiation, $P = \rho/3$, leading to $\rho \propto a^{-4}$. Finally, there is evidence that the expansion of the Universe is speeding up [27.37, 38]. The cause is unknown, but it is convenient to model it as *dark energy*, a perfect fluid with positive density but negative pressure. Postulating an equation of state of the form $P(t) = w(t)\rho(t)$ with $w(t) < 0$, one can once again solve (27.77) to obtain an expression for ρ as a function of the scale factor a.

Using this and the first Friedmann equation (27.75), the way the Universe evolves depending on its contents can be expressed through the *Hubble parameter* $H(a)$, defined as

$$H^2(a) \equiv \left(\frac{\dot a}{a}\right)^2 \tag{27.78}$$

$$= H_0^2\left[\Omega_\mathrm{M}a^{-3} + \Omega_\mathrm{R}a^{-4} + \Omega_k a^{-2} + \Omega_\mathrm{DE}\exp\left(3\int_0^a \frac{\mathrm{d}a'}{a'}\left[1 + w(a')\right]\right)\right], \tag{27.79}$$

where H_0 is the Hubble constant, which gives the expansion of the Universe at the current epoch, and the

dimensionless quantities Ω_{M}, Ω_{R}, Ω_k, and Ω_{DE} are the fractional contributions to the total density of, respectively, matter, radiation, spatial curvature, and dark energy

$$\Omega_{\mathrm{M}} = \frac{8\pi}{3H_0^2}\rho_{M,0}\,, \qquad \Omega_{\mathrm{R}} = \frac{8\pi}{3H_0^2}\rho_{\mathrm{R},0}\,,$$
$$\Omega_k = -\frac{k}{H_0^2}\,, \qquad \Omega_{\mathrm{DE}} = \frac{8\pi}{3H_0^2}\rho_{\mathrm{DE},0}\,, \tag{27.80}$$

with $\rho_{\mathrm{M},0}$, $\rho_{\mathrm{R},0}$, $\rho_{\mathrm{DE},0}$ the densities at the current epoch of matter, radiation, and dark energy, respectively.

An important task in cosmology is to determine

$$\boldsymbol{\Omega} \equiv (H_0, \Omega_{\mathrm{M}}, \Omega_{\mathrm{R}}, \Omega_k, \Omega_{\mathrm{DE}}, w(t))\,, \tag{27.81}$$

and especially to gain empirical insight into the enigmatic dark energy, by measuring its equation-of-state parameter $w(t)$. The main tools for studying the late-time evolution of the Universe are *standard candles.* These are distance markers for which both the redshift z and the luminosity distance D_{L} are known. For a source with intrinsic luminosity $\mathcal{L}$ and observed flux $\mathcal{F}$, D_{L} is defined through

$$\mathcal{F} = \frac{\mathcal{L}}{4\pi D_{\mathrm{L}}^2}\,. \tag{27.82}$$

If the Universe were Euclidean and never-changing, D_{L} would correspond to the familiar Euclidean notion of distance. However, due to the evolution of the Universe, D_{L} and z are related in a complicated way

$$D_{\mathrm{L}}(z) = c(1+z)\begin{cases} |k|^{-1/2}\sin\left(|k|^{1/2}\int_0^z \frac{\mathrm{d}z'}{H(z')}\right) & \text{for } \Omega_k < 0\,,\\ \int_0^z \frac{\mathrm{d}z'}{H(z')} & \text{for } \Omega_k = 0\,,\\ |k|^{-1/2}\sinh\left(|k|^{1/2}\int_0^z \frac{\mathrm{d}z'}{H(z')}\right) & \text{for } \Omega_k > 0, \end{cases} \tag{27.83}$$

where $H(z)$ is the Hubble parameter as a function of redshift. Since radiation will not be very important at late times, one can write (using $1/a = 1+z$)

$$H(z) = H_0\left[\Omega_{\mathrm{M}}(1+z)^3 + \Omega_k(1+z)^2 + (1-\Omega_{\mathrm{M}}-\Omega_k)E(z)\right]^{1/2}\,, \tag{27.84}$$

where $E(z)$ depends on the equation of state of dark energy. At late times, one can expand $w(t)$, or equivalently $w(z)$, as

$$w(z) = \frac{p_{\mathrm{DE}}}{\rho_{\mathrm{DE}}} = w_0 + w_a(1-a) + \mathcal{O}\left[(1-a)^2\right] \tag{27.85}$$

$$\simeq w_0 + w_a\frac{z}{1+z}, \tag{27.86}$$

in which case

$$E(z) = (1+z)^{3(1+w_0+w_a)}\mathrm{e}^{-3w_a z/(1+z)}\,. \tag{27.87}$$

From (27.83) and (27.84), it will be clear that given a large number of astrophysical sources for which the pairs (D_{L}, z) can be measured, one can constrain the parameters (27.81).

The most commonly used standard candles are Type Ia supernovae, whose luminosity is believed to be known within $\approx 10\%$ [27.37, 38]. However, this luminosity needs to be calibrated by comparison with different kinds of closer-by sources, leading to a *cosmic distance ladder*, each rung of which could contain unknown systematic errors. As pointed out by *Schutz* in 1986, GW signals from inspiraling neutron stars and black holes can provide an absolute measure of distance, with no dependence on other sources [27.28]. In the context of cosmology, they have been dubbed *standard sirens.* To see how this works, consider the amplitude of an inspiral signal as a function of time

$$\mathcal{A}(t) = \frac{1}{D_{\mathrm{L}}}\mathcal{M}^{5/3}g(\theta,\phi,\iota,\psi)\omega^{2/3}(t)\,. \tag{27.88}$$

Here $\mathcal{M} = (m_1 m_2)^{3/5}/(m_1+m_2)^{1/5}$ is the chirp mass, $g(\theta,\phi,\iota,\psi)$ is a known function of the sky position (θ,ϕ) and orientation of the orbital plane (ι,ψ), and $\omega(t) = \dot{\Phi}(t)$ is the instantaneous frequency. The chirp mass, and of course $\omega(t)$, can be obtained from the phase. Thus, if sky position and orientation are known, then from the amplitude one can infer the luminosity distance D_{L}.

27.5.2 Cosmography with Gravitational Wave Detectors

To make use of binary inspirals as standard sirens, what is needed is a way to obtain some information about redshift z, and also about the sky position (θ,ϕ) and orientation (ι,ψ) so that the luminosity distance D_{L} can be obtained from the GW amplitude. A variety of methods have been proposed to achieve this.

Using Electromagnetic Counterparts

Gamma ray bursts (GRBs) are among the most energetic electromagnetic events since the Big Bang. They roughly fall into two categories: short, hard GRBs and long, soft ones. It is believed (although only direct GW detection will provide a definitive answer) that short, hard GRBs are caused by the coalescence of two neutron stars, or a neutron star and a black hole [27.44]. Sometimes a GRB can be localized on the sky, providing (θ, ϕ). If this allows for the identification of the galaxy that was host to the inspiral event, then from its spectrum one can infer the redshift z. Finally, given a *network* of detectors, some information about the orientation (ι, ϕ) can also be obtained. Additionally, it is possible that GRBs are strongly beamed in a direction perpendicular to the inspiral plane, with inclination angle $\iota \lesssim 20°$.

In [27.29], *Nissanke* et al. investigated with what accuracy a network of advanced detectors would be able to do cosmology. We first note that with second-generation detectors, the maximum redshift out to which inspirals can be seen is $z \simeq 0.1$ for BNS and $z \simeq 0.2$ for NSBH. For small redshifts, the luminosity distance–redshift relationship, (27.83), reduces to

$$D_L \simeq \frac{cz}{H_0}, \tag{27.89}$$

which is just Hubble's law. This means that with advanced detectors, we will only be able to probe the Hubble constant H_0. However, since H_0 is an overall scaling factor in the full expression for D_L, its accurate and unbiased measurement is key to precision cosmology at the largest scales. Gravitational wave detection will provide us with a way of measuring H_0 without having to rely on any kind of cosmic distance ladder. Note that from (27.89), if redshift can be determined with essentially zero uncertainty, then the uncertainty ΔH_0 on the Hubble parameter is related to the distance uncertainty by

$$\frac{\Delta H_0}{H_0} = \frac{\Delta D_L}{D_L}. \tag{27.90}$$

This pertains to a *single* source; the accuracy will improve roughly as $\approx \sqrt{\mathcal{N}}$ for $\mathcal{N}$ events. *Nissanke* et al. found that with a network composed of the two Advanced LIGOs and Advanced Virgo, with $\mathcal{N} = 4$ BNS events one already has $\Delta H_0/H_0 \approx 13\%$, and with $\mathcal{N} = 15$, $\Delta H_0/H_0 \approx 5\%$ [27.29].

With Einstein Telescope it would be possible to see BNS events out to redshifts of several. In [27.32] and [27.33], detailed studies were made of how accurately the *full* set of cosmological parameters (27.81) could be measured. With $\mathcal{O}(1000)$ events with identifiable host galaxies over the course of 5–10 yr, Ω_M and Ω_{DE} could be constrained with an uncertainties comparable to what one finds in measurements of the cosmic microwave background (CMB). One can also use the CMB measurements for Ω_M, Ω_{DE}, and Ω_k, and their uncertainties, as priors, and focus on the dark energy equation of state parameter w, including its possible time dependence. Using a linear approximation to $w(z)$ as in (27.86), one can then compare accuracies in measuring w_0 and w_a, on the one hand using standard sirens seen by ET, and on the other hand considering the SNAP Type Ia supernova survey which may be available on the same timescale as ET. The results are shown in Fig. 27.7. The measurement quality is comparable in the two cases, but we stress once again that standard sirens allow for an *independent* measurement, with no need for a cosmic distance ladder.

Using a Prior on the Intrinsic Neutron Star Masses

Currently there are about 10 electromagnetically observed binary neutron star systems, with varying degrees of compactness. The distribution of neutron star masses in these binaries is relatively tight [27.108, 109],

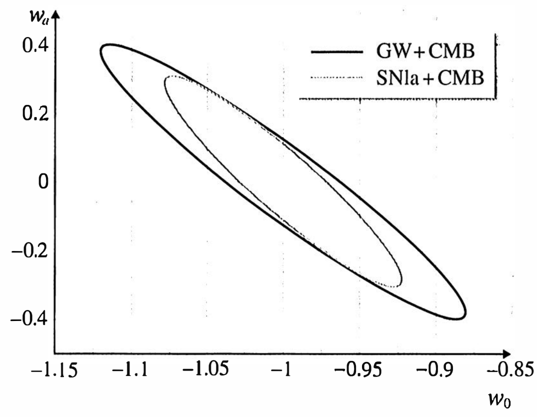


Fig. 27.7 Measurement uncertainties for the possible time dependence in the dark energy equation of state parameter w, modeled as $w(a) = w_0 + (1-a)w_a$, with a the scale factor (after [27.33]). The slightly larger, *brown ellipse* is for standard sirens as seen with ET, the *grey one* for the possible future SNAP Type Ia supernova survey. In both cases, prior information from the CMB is assumed for Ω_M, Ω_{DE}, and Ω_k

with mean $\mu_{NS} \simeq 1.34 M_\odot$ and standard deviation $\sigma_{NS} \simeq 0.06 M_\odot$. Now, the masses that are measured from a gravitational wave signal are not the physical ones m_{phys}, but the redshifted masses m_{obs}; for a source at redshift z, one has

$$m_{obs} = (1+z)m_{phys} . \quad (27.91)$$

As shown by *Taylor* et al., assuming an underlying, physical distribution of masses and comparing with the observed masses, one can obtain information about the redshifts of events without ever needing an electromagnetic counterpart [27.30]. With ≈ 100 BNS observations, a network of second-generation detectors would then allow the measurement of the Hubble constant with $\approx 10\%$ uncertainty.

Taylor and *Gair* applied this idea to a network of Einstein Telescopes, in which case one might have as many as 10^5 BNS signals per year [27.34]. Keeping H_0, Ω_M, Ω_{DE}, and Ω_k fixed, they also found that with this method, the dark energy equation of state parameters (w_0, w_a) in (27.86) could be measured with an accuracy comparable to the forecasted constraints from future SNIa surveys with CMB and other results as priors.

Using Galaxy Catalogs

Another exciting idea for measuring H_0 without the need for electromagnetic counterparts, and without having to restrict to a particular kind of inspiral event, was recently put forward by *Del Pozzo* [27.31]. He assumed three kinds of networks: the Advanced LIGO detectors together with Advanced Virgo (HLV), the same with the Japanese KAGRA added (HLVJ), and the five-detector network with IndIGO also included (HLVJI). Given an inspiral event, these networks will be able to localize it on the sky with different uncertainties; similarly, the distance $\hat{D}_L$ extracted from the gravitational wave signal will also be subject to errors. Combining these uncertainties, one obtains a large three-dimensional box in which the inspiral event could have occurred. A galaxy catalog will yield a list of potential host galaxies within this box, all having different redshifts $\hat{z}_i$. Using the Hubble law $D_L = cz/H_0$, the maximum-likelihood distance $\hat{D}_L$, together with this list of redshifts, leads to a list of possible values for the Hubble constant $\{H_{0,i}\}$. However, a *second* inspiral event will yield another list $\{H_{0,j}\}$, which will typically have only limited overlap with the first one. As more and more detections are made, the true value of the Hubble constant will quickly emerge. Del Pozzo cast this idea into the language of Bayesian analysis, and found that *after only 10 observations*, the 95% confidence level accuracy on H_0 is 14.5%, 7%, and 6.7% for the HLV, HLVJ, and HLVJI networks, respectively. After 50 observations, these numbers become 5%, 2%, and 1.8%, respectively; see also Fig. 27.8.

The advantage of this method is that it does not rely on any specific kind of source – in principle it can use *all* BNS, NSBH, and BBH detections. One issue will be the completeness of galaxy catalogs. However, it is possible to include in the analysis terms that describe the likelihood of observing a GW whose host galaxy was not detected by any survey because of its faintness; see [27.35].

MacLeod and *Hogan* explored a similar idea for measuring H_0 in the context of LISA, with GW signals from EMRIs, and using galaxy clustering [27.36]. Finally, a supermassive binary black hole at $z \lesssim 1$ would be sufficiently localizable with LISA that one might be able to find the host galaxy cluster, which would then yield a redshift, allowing for a measurement of the equation-of-state parameter of dark energy to within a few percent [27.110–112].

Using the Neutron Star Equation of State to Extract Redshift from the Gravitational Wave Signal

Recently, a method was developed to extract redshift information from the GW signal itself, at least in the case of BNS or NSBH. In the last stages of inspiral, a neutron star will be deformed and acquire a quadrupole moment Q_{ij} due to the tidal field $\mathcal{E}_{ij}$ of the compan-

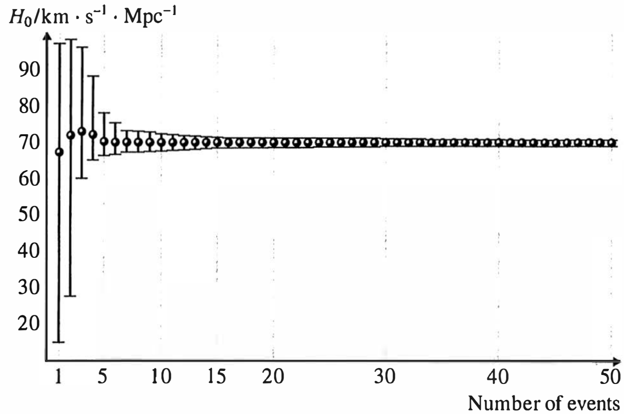


Fig. 27.8 Evolution of the medians and 95% confidence intervals for the Hubble constant as information from an increasing number of coalescence events is combined, using a galaxy catalog to obtain information on redshifts, for the HLVJI network of second-generation detectors (after [27.31])

ion object, and to leading order one can write $Q_{ij} = -\lambda \mathcal{E}_{ij}$. Here λ is the tidal deformability parameter, which depends both on the neutron star mass and the equation of state. These deformations have an influence on the orbital motion, which in turn gets imprinted onto the gravitational waveform. Such effects appear in the phase at 5 PN and 6 PN orders

$$\Phi(v) = \Phi_{\mathrm{PP}}(v) + \Phi_{\mathrm{tidal}}(v) , \tag{27.92}$$

where Φ_{PP} is the post-Newtonian phase under the assumption of point particles, and [27.79, 113]

$$\begin{aligned}\Phi_{\mathrm{tidal}} = \sum_{a=2}^{2} \frac{3\lambda_a}{128\eta M^5} &\times \left[-\frac{24}{\chi_a}\left(1 + \frac{11\eta}{\chi_a}\right)\left(\frac{v}{c}\right)^5 \right. \\ &- \frac{5}{28\chi_a}\left(3179 - 919\chi_a - 2286\chi_a^2 + 260\chi_a^3\right) \\ &\left. \times \left(\frac{v}{c}\right)^7 \right] ,\end{aligned} \tag{27.93}$$

where the sum is over the components of the binary, $\chi_a = m_a/M$, and $\lambda_a = \lambda(m_a)$, $a = 1, 2$. The function $\lambda(m)$ takes the form $\lambda(m) = (2/3)k_2 R^5(m)$, with k_2 the second Love number and $R(m)$ a neutron star's radius as a function of mass. Note that $\lambda(m)$ enters (27.93) only in the combination

$$\frac{\lambda(m)}{M^5} \propto \left(\frac{R}{M}\right)^5 \approx 10^5 . \tag{27.94}$$

Thus, although the tidal terms only appear at very high post-Newtonian order, they come with a large prefactor. Their effect will be noticeable already in advanced detectors, and certainly in Einstein Telescope.

Messenger and *Read* noted that the tidal contribution to the phase (27.93) only depends on *intrinsic* quantities [27.114]. Indeed, the expansion of the Universe as the signal travels from source to observer will cause the *observed* radius and mass to both be larger than the physical ones by a factor $(1 + z)$, which however will cancel from (27.94) and hence (27.93).

Einstein Telescope might see $\mathcal{O}(10^5)$ BNS sources. Some fraction of these could be used to determine the neutron star equation of state by measuring $\lambda(m)$. Once this is done, for each source in the other fraction one would be able to determine the *observed* masses $m_{\mathrm{obs}} = (1 + z)m_{\mathrm{phys}}$ from the low-order PN contributions to the phase, and the *intrinsic* masses m_{phys} from the tidal contribution (27.93). Hence both luminosity distances D_{L} and redshifts z can be inferred directly from the gravitational wave signals!

This will again allow for a fit of the luminosity distance–redshift relation $D_{\mathrm{L}}(z)$, thus constraining the cosmological parameters $\boldsymbol{\Omega}$ of (27.81), on condition that the uncertainties on redshift measurements are not too high. The latter depend on the equation of state, about which not much is currently known. Messenger and Read estimate that in the range $z = 0.1 - 1$, $\Delta z/z \approx 0.1$ for the *hardest* predicted equations of state (implying the greatest deformability), and $\Delta z/z \approx 0.4$ for *soft* equations of state. How this translates into constraints on the parameters $\boldsymbol{\Omega}$ is yet to be investigated.

27.6 Summary

In a few years' time, the second-generation gravitational wave detectors are due to deliver their first detections. This will herald a new era in the empirical study of gravitation. For the first time, we will have access to the genuinely strong-field dynamics of gravity. As a bonus, we will be able to look at weak-field gravity in a novel way, by searching for effects that may only show up at very large distance scales, such as the ones which gravitational waves must travel from source to observer.

In the older literature, studies of how well one might test GR with gravitational waves mostly took the form of estimates. With the advanced detector era approaching, the last few years have seen the development of hands-on data analysis pipelines to look for deviations from GR in actual detector data. Soon we will be searching for alternative polarization states, as well as for possible anomalies in the way that dynamical gravitational fields interact with themselves. For the latter, a full data analysis pipeline using coalescences of binary neutron stars is already in place. Binary black holes have a much richer dynamics, but the added complexity also makes for a formidable data analysis problem, the exploration of which has only just begun.

With third-generation ground-based detectors such as Einstein Telescope, and the space-based LISA, one

would be able to exploit not just the inspiral phase, but also the ringdown. As with the phasing coefficients, the Einstein equations impose relationships between characteristic frequencies and damping times, which effectively allow for a test of the no hair theorem. This will complement the test proposed earlier by Ryan, using extreme mass ratio inspirals.

Binary inspirals are *standard sirens* which can be used to probe the large scale structure of spacetime. Although the basic idea had already been proposed by Schutz as early as 1986, the last few years have seen the development of detailed methods, based on electromagnetic counterparts, exploiting the mass distribution of binary neutron stars, using galaxy catalogs, or employing knowledge of the neutron star equation of state.

The coming years are almost guaranteed to be a bonanza for gravitational physics. Either general relativity will be confirmed with more stringent tests than any that have been performed hitherto, or we will see deviations, which may well take the form of low-energy limits of quantum gravity effects. Any which way, the prospects are exciting indeed.

References

27.1 C.W. Misner, K.S. Thorne, J.A. Wheeler: *Gravitation* (Freeman, New York 1973)
27.2 C.M. Will: Living Rev. Relativ. **9**, 3 (2006)
27.3 J.H. Taylor, J.M. Weisberg: Astrophys. J. **253**, 908 (1982)
27.4 P.C.C. Freire, N. Wex, G. Esposito-Farèse, J.P.W. Verbiest, M. Bailes, B.A. Jacoby, M. Kramer, I.H. Stairs, J. Antoniadis, G.H. Janssen: The relativistic pulsar–white dwarf binary PSR J1738+0333 – II. The most stringent test of scalar-tensor gravity, Mon. Not. R. Astron. Soc. **423**, 3328–3343 (2012)
27.5 LIGO Scientific Collaboration, Virgo Collaboration: Predictions for the rates of compact binary coalescences observable by ground-based gravitational-wave detectors: Class, Quantum Gravity **27**, 173001 (2010)
27.6 G.M. Harry, LIGO Scientific Collaboration: Class. Quantum Gravity **27**, 084006 (2010)
27.7 The Virgo Collaboration: Advanced Virgo baseline design, VIR-027A-09 May 16 (2009), available online at https://tds.ego-gw.it/itf/tds/file.php?callFile=VIR-0027A-09.pdf
27.8 K. Kuroda, LCGT Collaboration: Class. Quantum Gravity **27**, 084004 (2010)
27.9 B.S. Sathyaprakash, LIGO Scientific Collaboration: Scientific Benefits of LIGO-India, LSC internal report G1100991 (2011)
27.10 H. Grote, LIGO Scientific Collaboration: Class. Quantum Gravity **25**, 114043 (2008)
27.11 H. Grote, LIGO Scientific Collaboration: Class. Quantum Gravity **27**, 084003 (2010)
27.12 Laser Interferometer Space Antenna (LISA), NASA; http://lisa.nasa.gov
27.13 N. Yunes, F. Pretorius: Phys. Rev. D **80**, 122003 (2009)
27.14 A. Nishizawa, A. Taruya, S. Kawamura: Phys. Rev. D **81**, 104043 (2010)
27.15 K. Chatziioannou, N. Yunes, N. Cornish: Phys. Rev. D **86**, 022004 (2012)
27.16 K. Hayama, A. Nishizawa: Model-independent test of gravity with a network of ground-based gravitational-wave detectors, Phys. Rev. D **87**, 062003 (2013)
27.17 L. Blanchet: Living Rev. Relativ. **9**, 4 (2006)
27.18 K.G. Arun, B.R. Iyer, M.S.S. Qusailah, B.S. Sathyaprakash: Class. Quantum Gravity **23**, L37 (2006)
27.19 K.G. Arun, B.R. Iyer, M.S.S. Qusailah, B.S. Sathyaprakash: Phys. Rev. D **74**, 024006 (2006)
27.20 C.K. Mishra, K.G. Arun, B.R. Iyer, B.S. Sathyaprakash: Phys. Rev. D **82**, 064010 (2010)
27.21 N. Cornish, L. Sampson, N. Yunes, F. Pretorius: Phys. Rev. D **84**, 062003 (2011)
27.22 T.G.F. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, J. Veitch, K. Grover, T. Sidery, R. Sturani, A. Vecchio: Phys. Rev. D **85**, 082003 (2012)
27.23 T.G.F. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, J. Veitch, K. Grover, T. Sidery, R. Sturani, A. Vecchio: J. Phys. Conf. Ser. **363**, 012028 (2012)
27.24 R.O. Hansen: J. Math. Phys. **15**, 2580 (1974)
27.25 F.D. Ryan: Phys. Rev. D **56**, 1845 (1997)
27.26 S. Gossan, J. Veitch, B.S. Sathyaprakash: Phys. Rev. D **85**, 124056 (2012)
27.27 I. Kamaretsos, M. Hannam, B.S. Sathyaprakash: Phys. Rev. Lett. **109**, 141102 (2012)
27.28 B.F. Schutz: Nature **323**, 310 (1986)
27.29 S. Nissanke, S.A. Hughes, D.E. Holz, N. Dalal, J.L. Sievers: Astrophys. J. **725**, 496 (2010)
27.30 S.R. Taylor, J.R. Gair, I. Mandel: Phys. Rev. D **85**, 023535 (2012)
27.31 W. Del Pozzo: Phys. Rev. D **86**, 043011 (2012)
27.32 B.S. Sathyaprakash, B.F. Schutz, C. Van Den Broeck: Class. Quantum Gravity **27**, 215006 (2010)
27.33 W. Zhao, C. Van Den Broeck, D. Baskaran, T.G.F. Li: Phys. Rev. D **83**, 023005 (2011)
27.34 S.R. Taylor, J.R. Gair: Phys. Rev. D **86**, 023502 (2012)
27.35 C. Messenger, J. Veitch: Avoiding selection bias in gravitational wave astronomy, New J. Phys. **15**, 053027 (2013)

27.36 C.L. MacLeod, C.J. Hogan: Phys. Rev. D **77**, 043512 (2008)
27.37 A.G. Riess, et al.: Astron. J. **116**, 1009 (1998)
27.38 S. Perlmutter, et al.: Astrophys. J. **517**, 565 (1999)
27.39 C. Fryer, K.C.B. New: Living Rev. Relativ. **6**, 2 (2006)
27.40 M. Maggiore, A. Nicolis: Phys. Rev. D **62**, 024004 (2000)
27.41 S. Capoziello, C. Corda: Int. J. Mod. Phys. D **15**, 1119 (2006)
27.42 E. Alesci, G. Montani: Int. J. Mod. Phys. D **14**, 1 (2005)
27.43 C. Charmousis, R. Gregory, N. Kaloper, A. Padilla: J. High Energy Phys. **10**, 066 (2006)
27.44 E. Nakar: Phys. Rep. **442**, 166 (2007)
27.45 Y. Gürsel, M. Tinto: Phys. Rev. D. **40**, 3884 (1989)
27.46 L. Blanchet, G. Faye, B.R. Iyer, B. Joguet: Phys. Rev. D **65**, 061501 (2002)
27.47 L. Blanchet, G. Faye, B.R. Iyer, B. Joguet: Erratum, Phys. Rev. D **71**, 129902 (2005)
27.48 L. Blanchet, B.S. Sathyaprakash: Class. Quantum Gravity **11**, 2807 (1994)
27.49 L. Blanchet, B.S. Sathyaprakash: Phys. Rev. Lett. **74**, 1067 (1995)
27.50 L.E. Kidder, C.M. Will, A.G. Wiseman: Phys. Rev. D **47**, R4183 (1993)
27.51 M. Maggiore: *Gravitational Waves. Volume 1: Theory and Experiments* (Oxford Univ. Press, Oxford 2007)
27.52 C.M. Will: Phys. Rev. D **57**, 2061 (1998)
27.53 C.M. Will, N. Yunes: Class. Quantum Gravity **21**, 4367 (2004)
27.54 E. Berti, A. Buonanno, C.M. Will: Class. Quantum Gravity **22**, S943 (2005)
27.55 A. Stavridis, C.M. Will: Phys. Rev. D **80**, 044002 (2009)
27.56 K.G. Arun, C.M. Will: Class. Quantum Gravity **26**, 155002 (2009)
27.57 D. Keppel, P. Ajith: Phys. Rev. D **82**, 122001 (2010)
27.58 W. Del Pozzo, J. Veitch, A. Vecchio: Phys. Rev. D **83**, 082002 (2011)
27.59 K. Yagi, T. Tanaka: Phys. Rev. D **81**, 064008 (2010)
27.60 C.M. Will: Phys. Rev. D **50**, 6058 (1994)
27.61 P.D. Scharre, C.M. Will: Phys. Rev. D **65**, 042002 (2002)
27.62 N. Yunes, F. Pretorius, D. Spergel: Phys. Rev. D **81**, 064018 (2010)
27.63 K. Yagi, N. Tanahashi, T. Tanaka: Phys. Rev. D **83**, 084036 (2011)
27.64 L.C. Stein, N. Yunes: Phys. Rev. D **83**, 064038 (2011)
27.65 K. Yagi, L.C. Stein, N. Yunes, T. Tanaka: Phys. Rev. D **85**, 064022 (2012)
27.66 K. Yagi, N. Yunes, T. Tanaka: Gravitational waves from quasi-circular black hole binaries in dynamical Chern–Simons gravity, Phys. Rev. Lett. **109**, 251105 (2012)
27.67 B. Zweibach: Phys. Lett. B **156**, 315 (1985)
27.68 A. Tseytlin: Phys. Lett. B **176**, 92 (1986)
27.69 S.H.S. Alexander, M.E. Peskin, M.M. Sheik-Jabbari: Phys. Rev. Lett. **96**, 081301 (2006)
27.70 S.H.S. Alexander, J. Gates, S. James: J. Cosmol. Astropart. Phys. **0606**, 018 (2006)
27.71 V. Taveras, N. Yunes: Phys. Rev. D **78**, 064070 (2008)
27.72 S. Mercuri, V. Taveras: Phys. Rev. D **80**, 104007 (2009)
27.73 S. Weinberg: Phys. Rev. D **77**, 123541 (2008)
27.74 C. Van Den Broeck, A.S. Sengupta: Class. Quantum Gravity **24**, 155 (2007)
27.75 C. Van Den Broeck, A.S. Sengupta: Class. Quantum Gravity **24**, 1089 (2007)
27.76 M. Vallisneri: Phys. Rev. D **86**, 082001 (2012)
27.77 E.T. Jaynes: *Probability Theory* (Cambridge Univ. Press, Cambridge 2003)
27.78 R. O'Shaughnessy, C. Kim, V. Kalogera, K. Belczynski: Astrophys. J. **672**, 479 (2008)
27.79 T. Hinderer, B.D. Lackey, R.N. Lang, J.S. Read: Phys. Rev. D **81**, 123016 (2010)
27.80 A. Buonanno, B.R. Iyer, E. Ochsner, Y. Pan, B.S. Sathyaprakash: Phys. Rev. D **80**, 084043 (2009)
27.81 R. O'Shaughnessy, J. Kaplan, V. Kalogera, K. Belczynski: Astrophys. J. **632**, 1035 (2005)
27.82 T.A. Apostolatos, C. Cutler, G.J. Sussman, K. Thorne: Phys. Rev. D **49**, 49 (1994)
27.83 L.E. Kidder: Phys. Rev. D **52**, 821 (1995)
27.84 P. Grandclément, J. Novak: Liv. Rev. Relativity **12**, 1 (2009)
27.85 T. Damour, A. Nagar: Fundam. Theor. Phys. **162**, 211 (2011)
27.86 A. Taracchini, Y. Pan, A. Buonanno, E. Barausse, M. Boyle, T. Chu, G. Lovelace, H.P. Pfeiffer, M.A. Scheel: Phys. Rev. D **86**, 024011 (2012)
27.87 P. Ajith, S. Babak, Y. Chen, M. Hewitson, B. Krishnan, J. T Whelan, B. Brügmann, P. Diener, J. Gonzalez, M. Hannam, S. Husa, M. Koppitz, D. Pollney, L. Rezzolla, L. Santamaría, A.M. Sintes, U. Sperhake, J. Thornburg: Class. Quantum Gravity **24**, S689 (2007)
27.88 L. Santamaría, F. Ohme, P. Ajith, B. Brügmann, N. Dorband, M. Hannam, S. Husa, P. Mösta, D. Pollney, C. Reisswig, E.L. Robinson, J. Seiler, B. Krishnan: Phys. Rev. D **82**, 064016 (2010)
27.89 P. Ajith, M. Hannam, S. Husa, Y. Chen, B. Brügmann, N. Dorband, D. Müller, F. Ohme, D. Pollney, C. Reisswig, L. Santamaría, J. Seiler: Phys. Rev. Lett. **106**, 241101 (2011)
27.90 R. Sturani, S. Fischetti, L. Cadonati, G.M. Guidi, J. Healy, D. Shoemaker: J. Phys. Conf. Ser. **243**, 012007 (2010)
27.91 R. Sturani, S. Fischetti, L. Cadonati, G.M. Guidi, J. Healy, D. Shoemaker, A. Vicerè: Phenomenological gravitational waveforms from spinning coalescing binaries (2011), arXiv:1012.5172 [gr-qc]
27.92 C. Vishveshwara: Nature **227**, 936 (1970)
27.93 E. Berti, J. Cardoso, V. Cardoso, M. Cavagliá: Phys. Rev. D **76**, 104044 (2007)
27.94 R. Ruffini, J.A. Wheeler: Phys. Today **24**, 30 (1971)
27.95 E.W. Leaver: Proc. R. Soc. **402**, 285 (1985)

27.96 I. Kamaretsos, M. Hannam, S. Husa, B.S. Sathyaprakash: Phys. Rev. D **85**, 024018 (2012)
27.97 E. Berti, V. Cardoso, C.M. Will: Phys. Rev. D **73**, 064030 (2006)
27.98 K.G. Arun, et al.: Class. Quantum Gravity **26**, 094027 (2009)
27.99 S.L. Liebling, C. Palenzuela: Living Rev. Relativ. **15**, 6 (2012)
27.100 P. Amaro-Seoane, L. Santamaria: Astrophys. J. **722**, 1197 (2010)
27.101 P. Amaro-Seoane: Stellar dynamics and extreme-mass ratio inspirals (2012), arXiv:1205.5240 [astro-ph.CO]
27.102 N.A. Collins, S.A. Hughes: Phys. Rev. D **69**, 124022 (2004)
27.103 K. Glampedakis, S. Babak: Class. Quantum Gravity **23**, 4167 (2006)
27.104 L. Barack, C. Cutler: Phys. Rev. D **75**, 042003 (2007)
27.105 S.J. Vigeland, S.A. Hughes: Phys. Rev. D **81**, 024030 (2010)
27.106 S. Vigeland, N. Yunes, L. Stein: Phys. Rev. D **83**, 104027 (2011)
27.107 C.L. Rodriguez, I. Mandel, J.R. Gair: Phys. Rev. D **85**, 062002 (2012)
27.108 B. Kiziltan, A. Kottas, S.E. Thorsett: The neutron star mass distribution, arXiv:1011.4291 [astro-ph]
27.109 R. Valentim, E. Rangel, J.E. Horvath: On the mass distribution of neutron stars, arXiv:1101.4872 [astro-ph]
27.110 K.G. Arun, B.R. Iyer, B.S. Sathyaprakash, S. Sinha, C. Van Den Broeck: Phys. Rev. D **76**, 104016 (2007)
27.111 K.G. Arun, C. Mishra, C. Van Den Broeck, B.R. Iyer, B.S. Sathyaprakash, S. Sinha: Class. Quantum Gravity **26**, 094021 (2009)
27.112 C. Van Den Broeck, M. Trias, B.S. Sathyaprakash, A.M. Sintes: Phys. Rev. D **81**, 124031 (2010)
27.113 F. Pannarale, L. Rezzolla, F. Ohme, J.S. Read: Phys. Rev. D **84**, 104017 (2011)
27.114 C. Messenger, J.S. Read: Phys. Rev. Lett. **108**, 091101 (2012)

Part E

General

Part E General Relativity and the Universe

28. Einstein's Equations, Cosmology, and Astrophysics

Paul S. Wesson

A compact, pedagogical review of our present understanding of the universe as based on general relativity is given. This includes the uniform models, with special reference to the cosmological constant; and the equations for spherically-symmetric systems, in a particularly convenient form that aids their application to astrophysics. New ideas in research are also outlined, notably involving extra dimensions.

28.1 Gravitation Today

General relativity is a remarkable subject, based on a few principles, yet covering a vast and intricate array of consequences. The present account is intended mainly as an overview; it is compact, yet reasonably complete. The material of Einstein's equations, cosmology, and astrophysics is treated in Sects. 28.2–28.4. Our understanding of these subjects has progressed so greatly that they have come to represent what might be termed academic industries. However, while many properties of the universe are well described by general relativity, there are indications that a deeper understanding may require an extended theory, the possible nature of which will be outlined in the conclusion of Sect. 28.5.

Einstein's theory has a broad literature, but those who work with it tend to gravitate to a few books (some of which are massive enough to justify the metaphor). The ones in the bibliography have different strengths, and together cover everything that is necessary for an understanding of the basics of the theory [28.1–5]. There are also certain subjects which occur in the following sections that are discussed in books and papers of a more technical sort [28.6–14]. Ideas at the forefront of research are of a diverse nature and are perhaps best approached through introductory accounts [28.15–18], since it is not clear where they will lead.

28.2 Einstein's Equations

This section is devoted to the genesis and properties of the field equations. The notation is standard, so $x^{0,123}$ are the coordinates of time and ordinary space. To avoid symbolic clutter, we adopt the usual ploy of imagining that we measure time, distance, and mass in units which make the speed of light c, Newton's constant of gravity G and Planck's constant of action h all equal to unity.

The so-called fundamental constants are, in fact, not very significant in their scientific content and are only constants in the sense of being useful conventions. They arise because the history of physics saw it useful to separate the things it deals with into categories, which in mechanics we label mass, length, and time [28.10, 11, 15–18]. We ascribe basic units for these things, denoted in the abstract by M, L, and T, and in practice by convenient measures like the gram, centimeter, and second. The latter are obviously man-made, but so are the former. The concepts of mass, length, and

time are instructive, and arise because of the ways in which humans perceive the world and comprehend it by the five senses. Over centuries of research, this approach has been honed, and nowadays we take it for granted that the equations of physics should be homogeneous in their physical dimensions.

Dimensional analysis – the traditional shortcut of the physicist – is really the application of an elementary form of group theory related to the Pi theorem. It provides a way of checking the dimensional consistency of the equations of physics under the permutations of the three base quantities M, L, and T. Dimensional analysis does not, of course, determine the dimensionless factors which may enter a problem, such as π or e. In this regard, it should be noted that the constants of physics *do* serve the useful purpose of converting a physical proportionality to an equation in numbers. To illustrate, let us consider the classical Kepler problem. In it, the Earth (mass m) orbits the Sun (mass M) with an azimuthal velocity (v) at a certain radial distance (r). The relative motion of the frames of reference of the two objects results in what historically has come to be called the centrifugal force mv^2/r. This is counterbalanced by the gravitational force of attraction between the objects, which following Newton we know to be proportional to the product of the masses and the inverse square of their separation. The essential physics of the Kepler orbit is described by the proportionality $mv^2/r \approx Mm/r^2$. However, to convert this to an *equation* we have to insert an appropriate constant G on the right-hand side. Its purpose is to transpose the physical characteristics of the quantities on the one side of the law to those on the other side, so it perforce has the physical dimensions of $M^{-1}L^3T^{-2}$. It is the somewhat arbitrary manner in which constants like G are introduced that has led several well-known workers to regard their presence in physics as accidental. By contrast, the canceling of the m on the left-hand side of the previous relation with the m on the right-hand side is *not* trivial. It is a consequence of Einstein's equivalence principle, to which we will return below. The simplicity of the Kepler problem, and particularly of the answer $v = \sqrt{GM/r}$, is due to this principle. Indeed, it is the fact that its laws are independent of the mass of a test object which makes gravitation a relatively simple science.

Quantum mechanics, in distinction to gravitation, is characterized by the unit of action h introduced by Planck and named after him. Both branches of science make use of c, the speed of light in vacuum. The complete suite of constants with their physical dimensions is thus $G = M^{-1}L^3T^{-2}$, $h = ML^2T^{-1}$, and $c = LT^{-1}$. While these constants are commonplace, it is important to realize that their dimensional contents do not *overlap*; each may be set to unity by an appropriate choice of units independent of the others. A corollary of this is that the mass of an object m can be geometrized, if so desired, in both subjects. The appropriate lengths are Gm/c^2 and h/mc, the Schwarzschild radius, and the Compton wavelength. The existence of these implies that it is possible, at least in principle, to construct a unified theory of gravitation and the interactions of particle physics which is based on geometry. It is also possible, as realized long ago by Planck and others, to use G, h, and c to define a *natural* set of units. (It is currently more common to use angular frequency than straight frequency in atomic problems, so $\hbar \equiv h/2\pi$ is the preferred unit.) The correspondence between natural or Planck units and the conventional gram, centimeter, and second can be summarized as follows:

$$
\begin{aligned}
1m_\mathrm{p} &\equiv \left(\frac{\hbar c}{G}\right)^{1/2} = 2.2\times10^{-5}\,\mathrm{g}\\
1\,\mathrm{g} &= 4.6\times10^{4}m_\mathrm{p}\\
1l_\mathrm{p} &\equiv \left(\frac{G\hbar}{c^3}\right)^{1/2} = 1.6\times10^{-33}\,\mathrm{cm}\\
1\,\mathrm{cm} &= 6.3\times10^{32}l_\mathrm{p}\\
1t_\mathrm{p} &\equiv \left(\frac{G\hbar}{c^5}\right)^{1/2} = 5.4\times10^{-44}\,\mathrm{s}\\
1\,\mathrm{s} &= 1.9\times10^{43}t_\mathrm{p}\,.
\end{aligned}
$$

In Plank units, all the constants G, $\hbar$, and c become unity and they consequently disappear from the equations of physics.

In general relativity, the masses of objects are nearly always taken to be constants. It is, therefore, a theory of accelerations rather than forces. The equivalence principle, noted above, thus states that test masses accelerate in a gravitational field at the same rate, irrespective of their composition. This refers not only to chemical composition, but also to contributions to effective mass from binding energy and electromagnetic and other types of energy. For a particle, the equivalence principle removes the distinction which might be made between the gravitational mass (the quantity concerned in the object's gravitational field) and the inertial mass (the quantity which measures the object's energy content). For a fluid, however, it will be seen below that this distinction still exists and indeed follows from the field equations. The latter should not, of course, lead to

consequences that depend on our choice of coordinates. The principle of covariance makes this arbitrariness of coordinate formal and by use of tensors ensures that the theory leads to results whose content is independent of how we describe things. As in other theories, in general relativity the prime objective is often the calculation of the path of a test particle. The geodesic principle provides a formal scheme for doing this. The analog of the distance between two nearby points in the four dimensions of spacetime is the elemental interval $\mathrm{d}s$, which also defines proper time. The interval can be extremized by varying it to isolate the shortest route, as in the symbolic relation $\delta[\int \mathrm{d}s] = 0$. The result is the geodesic equation, whose four components give the equations of motion along the time and spatial axes (the time component involves the energy while the components in ordinary 3-D (three-dimensional) space involve the momenta of the test particle). The three principles outlined in this paragraph, to do with equivalence, covariance, and the geodesic, form the basis of a theory which is both monolithic and far-reaching.

Einstein's field equations are usually presented as a match between the gravitational field and its source in matter. Some of the philosophical implications of this are still under discussion (see below), but the mathematical structure of the theory is straightforward. The interval between two nearby points in spacetime is defined via an extension of Pythagoras' theorem by $\mathrm{d}s^2 = g_{\alpha\beta}\,\mathrm{d}x^\alpha\,\mathrm{d}x^\beta$, where a repeated index upstairs and downstairs is shorthand for summation over time (x^0) and space (x^{123}). The metric tensor $g_{\alpha\beta}$ is a 4×4 array of potentials, which is taken to be symmetric and so has 10 independent elements. Generally the potentials depend on space and time $g_{\alpha\beta}(x^\gamma)$, but locally they are constants whose magnitudes may be set to unity, defining flat Minkowski spacetime where the diagonal components are $\eta_{\alpha\beta} = (+1, -1, -1, -1)$. The derivatives of $g_{\alpha\beta}$ with respect to the coordinates define the useful objects named after Christoffel, $\Gamma^\alpha_{\beta\gamma} \equiv (g^{\alpha\delta}/2)(g_{\beta\delta,\gamma} + g_{\gamma\delta,\beta} - g_{\beta\gamma,\delta})$. Here the partial derivative is denoted by a comma and should not be confused with the semicolon used to denote the covariant derivative, which takes into account the curvature of spacetime. (The covariant derivative of a vector, for example, is given by $V_{\alpha;\beta} = V_{\alpha,\beta} - \Gamma^\gamma_{\alpha\beta}A_\gamma$.) The Christoffel symbols figure in the geodesic equation mentioned above, which gives the acceleration of a test particle in terms of its four-velocity $u^\alpha \equiv \mathrm{d}x^\alpha/\mathrm{d}s$, via $\mathrm{d}u^\gamma/\mathrm{d}s + \Gamma^\gamma_{\alpha\beta}u^\alpha u^\beta = 0$. They are also used to define the Riemann tensor $R^\alpha_{\beta\gamma\delta}$, which may be shown to encode all of the relevant information about the gravitational field. However, the Riemann tensor has 20 independent components, whereas to obtain field equations to solve for the 10 elements of the metric tensor $g_{\alpha\beta}$ requires an object with the same number of components. This is provided by setting the upper index in $R^\alpha_{\beta\gamma\delta}$ equal to one of the lower indices, and summing, a process which produces the contracted tensor $R_{\mu\nu}$ named after Ricci. When this is again contracted by taking its product with the metric tensor in its upstairs or contravariant form, the result is $R = g^{\mu\nu}R_{\mu\nu} = R^0_0 + R^1_1 + R^2_2 + R^3_3$, the Ricci or curvature scalar. It can be thought of as a kind of measure of the average intensity of the gravitational field at a point in spacetime. Lastly, the combination $G_{\mu\nu} \equiv R_{\mu\nu} - (R/2)g_{\mu\nu}$ is of special interest because its 4-D (four-dimensional) covariant divergence is zero by construction: $G^\mu_{\nu;\mu} = 0$. The geometrical object $G_{\mu\nu}$ is known as the Einstein tensor and comprises the left-hand side of the field equations.

The preceding paragraph is standard material and familiar to many workers. However, it is not so widely known that Einstein wished to follow the same procedure for the *other* side of his field equations. That is, he wished to replace the common properties of matter, such as the density ρ and pressure p, by algebraic expressions. He termed the former *base wood* and the latter *fine marble*. In his later years, Einstein attempted to find a way to affect this transmutation by using an extra dimension. This had already been shown by Kaluza to unify the gravitational and electromagnetic equations of classical theory, and Kaluza suggested an extension to quantum theory that was designed to explain the magnitude of the electron charge in terms of the momentum in the fifth dimension. Unfortunately, to make algebraic progress, Kaluza was obliged to assume that the 5-D (five-dimensional) theory had functions independent of the fifth coordinate (the *cylinder* condition), and Klein took the extra dimension to be rolled up to an unobservably small size (*compactification*). These two conditions proved to be a mathematical straightjacket for the theory, which robbed it of much of its physical power and doomed Einstein's dream of a purely geometric account of gravity and matter. It was not until 1992 that a fully general 5-D theory was formulated, which explained matter as being induced in 4-D by the fifth dimension. Actually, it was devised by workers trying to find a geometric rationale for rest mass, who were originally ignorant of Einstein's forgotten dream. Later, however, the rediscovery of an old embedding theory of differential geometry due to Campbell showed that the 5-D theory (based on the 5-D Ricci tensor R_{AB}) contained the 4-D one (based on the 4-D Einstein ten-

sor $G_{\alpha\beta}$). This approach, known as space-time-matter theory, was joined in 1998 by the similar membrane theory, and it is now acknowledged that matter can be explained in geometric terms if so desired.

General relativity, in its regular 4-D form, matches the Einstein tensor $G_{\mu\nu}$ to an object which contains the phenomenological properties of matter, the energy-momentum tensor $T_{\mu\nu}$. The form of this depends somewhat on the type of matter involved, but the latter is commonly assumed to be a perfect fluid (with an isotropic pressure and a unique density and no viscosity). Then the appropriate matter tensor may be written $T_{\mu\nu} = (\rho + p)u_\mu u_\nu - p g_{\mu\nu}$, where u_μ are the four-velocities defined before. This form may look contrived, but it can be shown that the divergence $T^{\mu}_{\nu;\mu} = 0$ gives back the standard equations of motion in ordinary 3-D space plus the equation of continuity (conservation of mass) for the fluid.

Before joining the parts of Einstein's equations which describe the gravitational field ($G_{\mu\nu}$) and the matter ($T_{\mu\nu}$), it is necessary to tackle the notorious problem posed by the cosmological constant Λ. The mathematical possibility of adding a term $\Lambda g_{\alpha\beta}$ to the field equations arises because the metric tensor acts like a constant under covariant differentiation ($g_{\alpha\beta;\gamma} = 0$). The presence of such a term does not, therefore, upset the physical considerations used to identify the left-hand side ($G_{\mu\nu}$) and the right-hand side ($T_{\mu\nu}$) of the proposed field equations. Notwithstanding this, it *does* have physical consequences. Notably, in a 3-D spherically-symmetric distribution of matter, an acceleration appears which at radius r is $\Lambda r/3$. This is a repulsion for $\Lambda > 0$, but an attraction that augments gravity if $\Lambda < 0$. Einstein strongly disliked such a Λ term, because it acts on matter without being itself connected with matter. But Eddington, his contemporary, regarded the Λ term as the essential foundation of cosmology, and present observations do indeed indicate its importance (see later). There has been much wrangling about the cosmological constant, both in physics and philosophy. It continues to be a subject of controversy, because certain models of elementary particles imply intense vacuum fields which correspond to a large magnitude for Λ, in apparent contradiction with astrophysical observations which imply a small, positive value for Λ of order $10^{-56}\,\mathrm{cm}^{-2}$. The apparent discrepancy lies in the range 10^{80}–10^{120}. One reasonable way of explaining this is in terms of a 5-D theory, where Λ varies with scale depending on the size of the extra dimension (which, though, is controversial). A new angle on Λ may actually be gained by taking from particle physics the idea that the vacuum is not merely emptiness but the seat of significant physics, and joining this to the structure necessary for a tensor-based description of gravity like general relativity. The result is that the cosmological constant may be regarded as measuring the density and pressure of the vacuum, its equation of state being $\rho_v = -p_v = +\Lambda/8\pi$. This is neat, but not without its pitfalls. For example, it is common to take the physical dimension of Λ as L^{-2}, so with conventional units restored the dimensionally-correct form of the density is $\rho_\nu = \Lambda c^2/8\pi G$. This gives the impression that the vacuum is ultimately related to the strength of gravity, as measured by G. However, this is mistaken. Firstly, because there is a coupling constant $8\pi G/c^2$ in front of the energy-momentum tensor if the field equations are set up using conventional units, and this exactly cancels the similar factor in ρ_ν as written above. Secondly, the so-called fundamental constants are, in fact, disposable, as we saw before; and while it may be convenient to put them back at the end of a complicated calculation, the numerical size of a given constant depends on an arbitrary choice of units and has no real significance. By contrast, the geometrical factor in $\rho_\nu = \Lambda/8\pi$ does have some significance. It is composed of a conventional factor 2 connected with the standard way of expressing potentials and a factor 4π. This is connected with the fact that the surface area of a sphere of radius r around a given center in flat space is $4\pi r^2$, so the intensity of a conserved field necessarily falls off as $1/4\pi r^2$, and it is necessary to integrate over the same surface area in order to evaluate the strength of a source. This situation is identical to the one in classical electromagnetism as described by Maxwell's equations. Those equations are vector in nature and admit of a gauge term which is the gradient of a scalar function. Similarly, while Einstein's equations are tensor in nature, they too admit of a kind of gauge term. This is just the $\Lambda g_{\mu\nu}$ discussed above. In other words, the most satisfactory way to regard the cosmological constant is in terms of a kind of gauge term for the equations of general relativity.

Putting Einstein's field equations together is now – with the knowledge of the previous discussion – a simple business. We choose to keep the Λ term explicit, for mathematical generality and because it is indicated by modern observations. (Although it was skimped in certain older books, including a black-covered one published in 1973 that was big in size and influence.) The equations in standard form read

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi T_{\mu\nu} \,. \tag{28.1}$$

These equations, despite occupying only one line, entail a vast amount of physics. They are also remarkable in that they attempt to explain reality (as expressed by $T_{\mu\nu}$) in terms of a purely abstract quantity based on geometry (namely $G_{\mu\nu}$). While a precursor may be found in Maxwell's theory, Einstein's theory represents a fundamental break with older, mechanical ways of viewing the world. It is not the purpose of the present account to go into the many observations and tests which support the validity of the equations (28.1). However, given the abstract mode of their genesis, it is truly remarkable that they work.

28.3 Cosmology

In this section, the usual viewpoint is adopted that the universe started in some event like a big bang. This is indicated by the traditional evidence to do with the expansion, the microwave background, and nucleosynthesis.

To these should be added the more recent evidence of the integrated radiation produced by sources like stars in galaxies [28.7, 8]. These produce a background field dependent on astrophysical processes, which should not be confused with the cooled-down fireball radiation now seen as the 3K background. The integrated *light* from galaxies has a very low intensity, which in the optical band has only recently been constrained in a meaningful way. It is controlled by the intensity of the sources, the redshift effect of the Hubble expansion, and the age of the universe. The last factor is important, and models of the integrated radiation from galaxies confirm that the present age is $t_0 \simeq 13\times10^9$ yr approximately. The night sky is dark because the universe has a finite past history, as expected if there was something like a big bang.

The present universe, on the basis of supernova and other data, appears to be accelerating under the influence of the cosmological constant or some similar scalar field. There is also ample evidence from the structure of spiral galaxies, the morphology of clusters of galaxies, and the gravitational lensing of distant sources like quasars that there is a significant density of dark matter in the universe. The nature of this is controversial, but it could be elementary particles of some kind with a low effective temperature. Ordinary matter, of the kind seen in stars and the optical parts of galaxies, makes up a relatively small fraction of the whole, especially in comparison to the effect of the cosmological constant regarded as a density for the vacuum (Sect. 28.2). The relative densities of the vacuum, dark matter, and ordinary matter are 74% : 22% : 4% approximately. We see that the stuff of traditional astronomy is a mere smattering.

It is difficult to match the aforementioned data to any simple model of cosmology. It is particularly difficult to find a single set of parameters which gives the evolution of the scale factor $R(t)$ as a function of cosmic time, notably in regard to the supernova data indicating acceleration at the present epoch. For this reason, the current picture is largely qualitative: following the big bang, there appears to have been a phase of rapid or inflationary expansion, with the equation of state of the vacuum ($p = -\rho$), when the universe became relatively smooth; then there was a hot period when the equation of state of the matter was close to that of radiation ($p = \rho/3$); and this evolved with cooling into the later phase we observe at present, when the matter is cold and behaves like dust ($p \simeq 0$), but where the Λ-like expansion is still dominant. To model these different phases, we need to take the field equations (28.1) and find relevant solutions.

The required solutions are named after Friedmann, Robertson, and Walker (FRW). The first reduced the field equations to a pair of convenient relations which will be examined below. The latter two workers isolated the corresponding form of the interval, which is useful for calculating distances and related quantities. The 4-D interval consists of two parts: a simple time and a measure for the 3-D distance whose form ensures that all places are equivalent. The Robertson–Walker interval is given by

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \frac{R^2(t)}{\left(1 + kr^2/4\right)^2}\left[\mathrm{d}r^2 + r^2\,\mathrm{d}\Omega^2\right] . \qquad (28.2)$$

Here $\mathrm{d}\Omega^2 \equiv (\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)$ defines the angular part of the metric in spherical polar coordinates. The radial part is expressed for ease in terms of a measure that is chosen to be comoving with the matter, which means that r in (28.2) is merely a distance *label*, the same at all time for a given galaxy. The *actual* (changing) distance involves the scale factor $R(t)$, which measures

the separation between two typical galaxies at time t. The rate of expansion is given by Hubble's parameter $H \equiv \dot{R}/R$, where an overdot denotes the total derivative with respect to time. The second derivative of $R(t)$ is measured for historical reasons by the deceleration parameter, $q \equiv -\ddot{R}R/\dot{R}^2$. This is dimensionless, while H has the units of an inverse time. (The present value of H is about 70 km/s/*Mpc* in terms of its traditional but rather perverse unit, and galaxies that are not too distant recede at velocities proportional to this and the distance.) The constant k in (28.2) is a normalized measure of the curvature of ordinary 3-D space, and can be positive, negative or zero (see below). It should be noted that an alternative form of (28.2) appears in some texts, obtained from it by a change in the radial coordinate, thus

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - R^2(t)\left(\frac{\mathrm{d}r^2}{(1-kr^2)} + r^2\,\mathrm{d}\Omega^2\right). \quad (28.3)$$

This is useful if we choose to measure r from ourselves considered as *center*, whereas (28.2) is spatially isotropic and provides a more *global* measure. Of course, for both forms, there is no real center and no boundary.

When the Robertson–Walker interval is used in conjunction with the Einstein field equations (28.1), the latter take the form of two relations which were studied by Friedmann. The assumption that the density ρ and pressure p of the cosmological fluid are isotropic and homogeneous (= uniform) causes the partial differential equations (28.1) to become ordinary differential equations in the scale factor $R(t)$ which measures the expansion. Friedmann's equations are

$$8\pi\rho = \frac{3k}{R^2} + \frac{3\dot{R}^2}{R^2} - \Lambda \quad (28.4)$$

$$8\pi p = \frac{-k}{R^2} - \frac{\dot{R}^2}{R^2} - \frac{2\ddot{R}}{R} + \Lambda\,. \quad (28.5)$$

Here the constant k, as mentioned above, measures the curvature of the 3-D ordinary space of the models and is normalized to have the values ± 1, 0. It can be thought of as indicating the relative contributions of the kinetic energy and gravitational binding energy for a unit volume of the fluid. In the absence of Λ, $k=-1$ means that the balance of energies is in the direction of continued expansion, $k=+1$ means that the fluid eventually stops expanding and collapses under its own gravity, while $k=0$ means an exact balance with a continuing but slowing expansion. However, Λ is not absent in the real universe, which considerably complicates the dynamical solutions of (28.4) and (28.5), most of which can only be isolated by numerical means.

Some instructive things emerge from the two Friedmann equations (28.4) and (28.5) when they are combined in appropriate ways. For this, it is useful to replace Λ by its equivalent vacuum properties (see above) and write the total density and pressure as $\rho = \rho_m + \rho_v$, $p = p_m + p_v$ with matter and vacuum parts. Then combining (28.4) with three times (28.5) to eliminate k gives

$$\ddot{R} = \frac{-(4/3)\pi R^3(\rho + 3p)}{R^2}\,. \quad (28.6)$$

This is seen to be a quasi-Newtonian law of inverse-square attraction, when we recall that the physical distance in 3-D is proportional at any time to the scale-factor $R(t)$, although this symbol does not imply a physical boundary since the cosmological fluid is continuous. It is noteworthy that the effective gravitational mass of a portion of the fluid is proportional to the combination $(\rho + 3p)$, not the Newtonian ρ (which is only recovered for $p \ll \rho$). Accordingly, the combination $(\rho + 3p)$ is called the gravitational density. For pure vacuum, this combination is negative for $\Lambda > 0$ since $p = -\rho = -\Lambda/8\pi$, and this is why a universe dominated by a positive cosmological constant experiences a cosmic repulsion. Another instructive thing emerges when the first derivative of (28.4) is combined with (28.5) to eliminate $\ddot{R}$, to give

$$\dot{\rho} = -(\rho + p)\left(\frac{3\dot{R}}{R}\right). \quad (28.7)$$

This is seen to be a kind of stability relation for the universe, in the sense that the density adjusts in proportion to the expansion rate and the combination $(\rho + p)$. This is not a gravitational effect, and accordingly the noted combination is called the inertial density. For pure vacuum, the combination $(\rho + p)$ is zero since the equation of state is $p = -\rho$. So the vacuum has constant density (and pressure) even though the matter in the universe is expanding.

It is apparent from the above that the universe according to Einstein can have properties quite different from those predicted by Newton. The reasons for this have primarily to do with the cosmological constant, the possibility that the pressure of matter may be a significant fraction of the energy density, and the fact that the speed of light is large but finite. The last of these has consequences which are subtle but ubiquitous. To briefly review these, let us temporarily reinstate conventional (nongeometrical) units for c. Then it is obvious that as we look to greater distances we also look back in time. Advances in observational techniques are such

that we can soon expect to be able to study in detail the first generation of galaxies. At greater distances we would ‚see' the primordial plasma from which the galaxies formed, and beyond that would be the zipping sea of strange particles being carried along by inflation. Since the universe is isotropic about every point and about *us*, we might in principle be able to ‚see' the big-bang fireball, which would resemble a glowing shell all around us.

Horizons, however, might block our view of the remote cosmos as they do our view of the distant parts of the Earth [28.1–5]. In the cosmological context, there are actually two kinds of horizon: an event horizon separates those galaxies we can see from those we cannot ever see even as $t \to \infty$; while a particle horizon separates those galaxies we can see from those we cannot see now at $t = t_0 \simeq 13 \times 10^9$ yr. FRW models exist which have both kinds of horizon, one but not the other, or neither. To investigate these, consider the path of a photon which moves radially through a universe where distance is defined by the Robertson–Walker metric. We put $\mathrm{d}s = 0$, $\mathrm{d}\theta = \mathrm{d}\phi = 0$ in (28.3) and obtain the (coordinate-based) velocity as $\mathrm{d}r/\mathrm{d}t = \pm c(1 - kr^2)^{1/2}/R(t)$. The sign choice here corresponds to whether the photon is moving towards or away from us. More importantly, we see that the *speed* of the photon is *not* just c. It actually depends on $R(t)$, which is given by the Friedmann equations (28.4) and (28.5). This means that the distance to the particle horizon, which defines that part of the universe in causal communication with us, can be quite complicated to work out. However, algebraic expressions can be written down for the simple case where $\Lambda = 0$ and $p = 0$. Then for the three values of the curvature constant, the distances are given by

$$\begin{aligned} d &= \frac{c}{H_0(2q_0 - 1)^{1/2}} \cos^{-1}\left(\frac{1}{q_0} - 1\right), \\ k &= +1, q_0 > \frac{1}{2} \\ d &= \frac{2c}{H_0} = 3ct_0 , \\ k &= 0, q_0 = \frac{1}{2} \\ d &= \frac{c}{H_0(1 - 2q_0)^{1/2}} \cosh^{-1}\left(\frac{1}{q_0} - 1\right), \\ k &= -1, q_0 < \frac{1}{2} . \end{aligned} \tag{28.8}$$

The Hubble parameter and deceleration parameter used here were defined above and are to be evaluated at the present epoch. It is apparent from these relations that the size of that part of the universe we can see is *not* just given by the product of the speed of light and the age.

The redshift z is in some ways a better parameter to use as a cosmological measure than either the distance or the time. It is a parameter which is directly observable, and it runs smoothly from us ($z = 0$), through the populations of galaxies and quasars ($z \simeq 1 - 10$), and in principle all the way to the big bang ($z \to \infty$). It is defined in terms of the scale factor of the Robertson–Walker metric at present (t_0) and at emission (t_e) by $1 + z \equiv R(t_0)/R(t_e)$. This neatly sidesteps long-running arguments about whether the redshift is *caused* by the Doppler effect, gravity, or some other agency, which are frame-dependent in general relativity and cannot be uniquely identified. The noted definition merely makes a statement about light waves and a ratio of scales. (It might even be imagined that the universe is momentarily static at the two instants which define the redshift, with no information available as to what happened in-between.) Notwithstanding the utility of the redshift as a measure, it is still true that most workers have a mental picture of a universe that evolves through stages separated in time. This is actually acceptable, provided that the epoch is used only in a relative sense, as an ordering device. Let us, therefore, return to this mode of organization and list the solutions of the Friedmann equations (28.4) and (28.5) relevant to the successive phases of the universe.

Inflation is characterized by a rapid expansion under the influence of the cosmological constant or some similar measure of vacuum energy. The appropriate solution of (28.4) and (28.5) was found by de Sitter in the early days of general relativity and is given by

$$p = -\rho = -\Lambda/8\pi, R(t) \approx \mathrm{e}^{t/L}, k = 0 . \tag{28.9}$$

The length scale here is related to the cosmological constant by $\Lambda = 3/L^2$ (the proportionality sign indicates that the scale factor is arbitrary up to a constant). The present universe also appears to have a significant value of Λ, which corresponds to a length L of order 10^{28} cm. The interval corresponding to (28.9) is

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \mathrm{e}^{2t/L}(\mathrm{d}r^2 + r^2\,\mathrm{d}\Omega^2) . \tag{28.10}$$

There is an alternative form of this cosmological metric, which is related by a coordinate transformation but is local in nature, thus

$$\mathrm{d}s^2 = \left(1 - \frac{\Lambda r^2}{3}\right)\mathrm{d}t^2 - \frac{\mathrm{d}r^2}{(1 - \Lambda r^2/3)} - r^2\,\mathrm{d}\Omega^2 . \tag{28.11}$$

This form of the de Sitter metric has been extensively used to model quantum-mechanical processes in the early universe, like tunneling. Such processes could be of great importance if it should be shown that general relativity needs to be extended in some way. For example, it then becomes feasible to explain the big bang as a quantum event, perhaps in a higher-dimensional manifold. In this regard, it can be mentioned that for both signs of Λ (28.11) can be embedded in a 5-D manifold which is *flat*, in which case (28.11) resembles a 4-D pseudosphere with radius L [28.2, 13]. Similarly, (28.10) can be embedded in 5-D Minkowski space.

Following inflation, the universe is believed to have passed through a hot period when the matter had an equation of state similar to that of radiation. A solution of the Friedmann equations (28.4) and (28.5) has been known for a long while that has the noted properties, although it was formulated before the importance of Λ was realized. The formal solution has

$$\begin{aligned} &p = \rho/3 = 1/32\pi t^2 , \quad R(t) \approx t^{1/2} , \\ &k = 0 , \qquad\qquad\qquad\quad \Lambda = 0 . \end{aligned} \tag{28.12}$$

This solution needs to be modified as regards its global properties for $\Lambda \neq 0$, but its local properties are still those necessary for nucleosynthesis of the kind needed to explain the observed abundances of the elements.

Later, when the matter had cooled, the universe is believed to have evolved into a cold phase which persists to the present and which is characterized by a value for the matter pressure which is effectively zero. The formal solution of (28.4) and (28.5) has

$$\begin{aligned} &p = 0 , \quad \rho = 1/6\pi t^2 , R(t) \approx t^{2/3} \\ &k = 0 , \quad \Lambda = 0 . \end{aligned} \tag{28.13}$$

This solution, like the previous one, needs to be modified in regard to its global properties for $\Lambda \neq 0$. The solution (28.13) is named after Einstein/de Sitter and should not be confused with the straight de Sitter solution (28.9). For many years, (28.13) was considered to be the closest approximation to the real universe. It is slightly ironic that modern data indicate that the old solution (28.9), with the cosmological constant so detested by Einstein, may be closer to the truth.

28.4 Astrophysics

The application of general relativity to astrophysical systems is simpler than to cosmology for one main reason: the influence of the cosmological constant is negligible. For this reason, we largely ignore it in this section. Also, despite what was stated in Sect. 28.2 about the disposability of the so-called fundamental constants, G and c are now made explicit in order to bring out the comparison with Newtonian theory and special relativity.

Many astrophysical systems are approximately spherically symmetric in ordinary 3-D space. The solar system is like this, although as a solution of Einstein's equations (28.1) it is exceptionally simple because it is approximately empty of matter except for the Sun (mass M). The interval may be regarded as an extended version of the local de Sitter one (28.11) and is given by

$$\begin{aligned} \mathrm{d}s^2 = &\left(1 - \frac{2GM}{c^2 r} - \frac{\Lambda r^2}{3}\right) \mathrm{d}t^2 \\ &- \frac{\mathrm{d}r^2}{(1 - 2GM/c^2 r - \Lambda r^2/3)} - r^2\,\mathrm{d}\Omega^2 . \end{aligned} \tag{28.14}$$

This is the familiar form, but it should be noted that the potential can be written $2G(M + M_v)/c^2 r$, where $M_v = (4/3)\,\pi r^3 \rho_v$ is the effective mass of the vacuum due to its equivalent density $\rho_v = \Lambda c^2/8\pi G$ (Sect. 28.2). It should also be noted that while the local de Sitter solution (28.11) can be embedded in flat 5-D, the Schwarzschild–de Sitter solution (28.14) cannot be embedded in a flat space of less than six dimensions. The fact that (28.14) successfully accounts for the dynamics of the solar system and binary pulsars, thereby establishing the validity of general relativity, also means that any extra dimensions must play an insignificant role in much of astrophysics.

To study other astrophysical systems where there is substantial matter, we assume the latter to be a spherically-symmetric perfect fluid described by the scalars ρ and p for the density and pressure. It is convenient to take the interval in the form

$$\mathrm{d}s^2 = \mathrm{e}^{\sigma} c^2\,\mathrm{d}t^2 - \mathrm{e}^{\omega}\,\mathrm{d}r^2 - R^2\,\mathrm{d}\Omega^2 . \tag{28.15}$$

Here σ and ω are metric coefficients that in general depend on the time t and a radial measure r, which can be

chosen to be comoving with the matter [28.1–5]. The latter may flow either inwards or outwards, but an element of it then maintains the same radial label r (as in the Robertson–Walker metric of Sect. 28.3). By contrast, $R = R(t, r)$ is really another metric coefficient and measures the dynamics of the fluid, although in such a way that $2\pi R$ is the circumference of a great circle around the center of the distribution. With this setup, it is the inequality of r and R in (28.15) which characterizes the departure of ordinary 3-D space from flatness due to the gravitational field of the fluid.

Given the interval (28.15), the question arises of how to write Einstein's equations (28.1) in the most informative manner. In many texts, they are written as long strings of symbols relating the derivatives of the metric coefficients σ, ω, and R to the properties of matter ρ and p. For problems of the type being considered here, there will in general be four equations for the five unknowns. Therefore, one relation may be specified in order to balance things and hopefully find a solution (ways to do this are examined below). However, in such problems it is often useful to define a function which is first order in the derivatives as a *new* unknown and rewrite the four *second*-order partial differential equations as five *first*-order ones [28.9]. For the current problem, it was found some while ago by Podurets and Misner and Sharp that the appropriate new function to define is a measure of the mass of the fluid interior to radius r at time t, that is, $m = m(r, t)$. The upshot is a set of five first-order differential equations in three metric coefficients (σ, ω, and R) and three properties of matter (ρ, p, and m). Not only does this improve the tractability of the algebra, it also (after some manipulation) leads to a set of equations which have much greater physical meaning.

Writing the definition of the mass function as a relation with other quantities, the full set of field equations is

$$\frac{2Gm}{c^2R} = 1 + \frac{e^{-\sigma}}{c^2}\left(\frac{\partial R}{\partial t}\right)^2 - e^{-\omega}\left(\frac{\partial R}{\partial r}\right)^2 , \quad (28.16)$$

$$\frac{\partial m}{\partial t} = \frac{-4\pi pR^2}{c^2}\frac{\partial R}{\partial t} , \quad (28.17)$$

$$\frac{\partial m}{\partial r} = 4\pi\rho R^2\frac{\partial R}{\partial r} , \quad (28.18)$$

$$\frac{\partial \sigma}{\partial r} = \frac{-2}{p+\rho c^2}\frac{\partial p}{\partial r} , \quad (28.19)$$

$$\frac{\partial \omega}{\partial t} = \frac{-2c^2}{p+\rho c^2}\frac{\partial \rho}{\partial t} - \frac{4}{R}\frac{\partial R}{\partial t} . \quad (28.20)$$

Experience shows that the first of the equations (28.16) is usually the hardest to solve. However, it is helpful to note that it involves a balance between the Schwarzschild-like gravitational potential Gm/c^2R, the kinetic energy per unit mass of the fluid $(\partial R/\partial t)^2$, and a measure of the departure of ordinary space from flatness, or equivalently the binding energy per unit mass of the fluid stored in the gravitational field $(\partial R/\partial r)^2$. The second equation (28.17) is best interpreted from right to left. It says, loosely speaking, that the force due to the pressure p acting over a shell of area $4\pi R^2$ that moves at a velocity $\partial R/\partial t$ forms a quantity which in mechanics would be termed a rate of work or power, and that the mass of the fluid responds by changing at a rate consistent with Einstein's formula for the equivalent energy mc^2. The third equation (28.18) would on integration give the usual Newtonian expression for the mass of a portion of the fluid ($m = 4\pi R^3/3$) if the space were flat ($R = r$), but since it is not, (28.18) gives the corresponding differential form for the mass of the fluid as affected by its own gravitational field. The last two equations, (28.19) and (28.20), relate the metric coefficients to the properties of the matter responsible for curving spacetime.

Solving (28.16)–(28.20) can be achieved once an extra relation is specified which balances the number of equations and the number of unknowns. There are also numerous solutions in the literature which were found by more tedious means and whose physical meanings may be elucidated by employing (28.16)–(28.20). It would be redundant to list those solutions here, especially since reviews are available [28.1–6]. The relations (28.16)–(28.20) have been applied to a wide range of problems, since they cover everything from the global cosmological fluid (the Friedmann equations included) to tiny perturbations of it [28.9]. Thus, they lead to a more objective form of the cosmological principle, in which all intelligent observers judge the universe to be the same everywhere, not merely in terms of the density and pressure but in terms of *dimensionless* combinations of these and other parameters. While at the other end of the spectrum, they can be used to study the growth of material around a quantum seed to form a protogalaxy. The equations in the form (28.16)–(28.20) are especially useful in understanding the behavior of matter under extreme circumstances, such as when the pressure approaches the energy density and the velocity of sound approaches the speed of light. New solutions like this certainly await discovery.

Ways to specify a condition which makes the set of equations (28.16)–(28.20) determinate are also various, and some examples follow:

a) An equation of state, $p = p(\rho)$, is the traditional approach. This is particularly efficacious if information about the microscopic state of the matter is available, for example from spectral observations of a real system.
b) Boundary conditions, in the broad sense, can help to restrict the form of a solution. These may include continuity conditions on the metric tensor if there is a join to another solution; or physical conditions, such as ones on the pressure at the center and periphery of a system.
c) Morphological constraints, such as self-similarity. The latter technique is especially relevant to astrophysical systems, which often lack sharp boundaries or other scales. A distribution completely free of scales may be described by defining a dimensionless combined variable (say ct/r), so enabling the problem to be posed in ordinary rather than partial differential equations, which are easier to solve. A distribution with a single scale may be tackled using a refinement of this technique, so that problems like phase changes which involve a change in size of a physical parameter can be treated.

The preceding options are not exhaustive and in any case there is the alternative of numerical integration. However, due to the nonlinearity of Einstein's equations, an exact algebraic solution is especially valuable. The search for new solutions is left as an exercise for the motivated reader.

28.5 Conclusion

General relativity is in the happy situation of being agreed upon by the great majority of workers and being verified by observations that stretch from the solar system to the most remote quasars. Much of cosmology can be treated using the Friedmann equations (28.4) and (28.5) for a uniform fluid, and much of astrophysics can be handled by the more complicated equations (28.16)–(28.20) for a spherically-symmetric fluid. In these two areas, it remains to find a single model that describes the whole history of the universe and solutions that describe the diversity of its constituent systems. Notwithstanding these technical shortcomings, it is still true to say that Einstein's theory provides quite a good account of the real universe.

It is also true, however, that a shift in our understanding of the classical universe will occur if a way is found to unify it with the quantum theory of particle interactions. That a connection exists is already hinted by the cosmological-constant problem, wherein the energy density of the vacuum is observed to be small on macroscopic scales but inferred to be large on microscopic scales. This problem would, of course, disappear if the properties of the vacuum prove to be variable. However, even this compromise will entail significant changes to our current accounts of both cosmology and particle physics. In fact, most workers believe that new physics will inevitably emerge from a unification of our present classical and quantum theories. Currently, the preferred route to unification is via extra dimensions. The basic extension is to 5-D, which as mentioned before is commonly called space-time-matter theory or membrane theory, whose main concerns are with classical matter and particles, respectively. For cosmology, perhaps the main consequence of the fifth dimension is the realization that the 4-D big bang is a kind of artefact, produced by an unfortunate choice of coordinates in a flat 5-D manifold. More generally, 5-D relativity is a unified theory of gravitation, electromagnetism, and a scalar field. It is the classical analog of the quantum interactions of the spin-2 graviton, the spin-1 photon, and a spin-0 scaleron. The last may be related to the Higgs boson, which is believed to be responsible for the finite masses of other objects (although in the classical theory masses also involve shifts in spacetime along the fifth dimension). As to the old question: Why do we not *see* the extra dimension? Well, in a way we do: it is the mass/energy all around us. This may sound strange, but adding one or more extra dimensions is actually the most effective way to extend general relativity so as to obtain new physics without upsetting established knowledge.

Our understanding of even established theory could do with some improvement, especially in what might be termed the psychology of cosmology. Anyone who has taught cosmology knows that even bright students have difficulty with the concepts raised by Einstein's theory. Moreover, even some researchers have an inadequate idea of what the big bang must have been like. This

is largely because human beings are imprinted from childhood with everyday constructs which leave them ill-equipped as adults to visualize a universe without a center or a boundary. Yet if the density and pressure depend only on the time then logic tells us that neither thing can exist. Confusion is engendered by calling the big bang an explosion, because this brings to mind a conventional bomb that sends shrapnel out from a point in 3-D space until it hits some obstruction like a wall. Insofar as an analogy can be made, the big bang should be imagined as a kind of explosion that fills all of 3-D space at the same moment, as if there is an indefinitely large number of bombs which are wired together so that they all detonate at the same instant. Even this description does not get across all of the subtleties of the Einstein singularity, but hopefully more attention will be given in the future to thinking in the right way.

References

28.1 S. Carrol: *Spacetime and Geometry: An Introduction to General Relativity* (Addison-Wesley, San Francisco 2004)

28.2 W. Rindler: *Relativity: Special, General and Cosmological* (Oxford Univ. Press, Oxford 2001)

28.3 J.N. Islam: *An Introduction to Mathematical Cosmology* (Cambridge Univ. Press, Cambridge 1992)

28.4 C.W. Misner, K.S. Thorne, J.A. Wheeler: *Gravitation* (Freeman, San Francisco 1973)

28.5 H.P. Robertson, T.W. Noonan: *Relativity and Cosmology* (Saunders, Philadelphia 1968)

28.6 D. Kramer, H. Stephani, M. MacCallum, E. Herlt: *Exact Solutions of Einstein's Field Equations* (Cambridge Univ. Press, Cambridge 1980)

28.7 J.M. Overduin, P.S. Wesson: *The Light/Dark Universe* (World Scientific, Singapore 2008)

28.8 J.M. Overduin, P.S. Wesson: Dark matter and background light, Phys. Rep. **402**, 267–406 (2004)

28.9 P.S. Wesson: A new look at the cosmological principle, Astron. Astrophys. **68**, 131–137 (1978)

28.10 V. Petkov (Ed.): *Relativity and the Dimensionality of the World* (Springer, Dordrecht 2007)

28.11 V. Petkov (Ed.): *Minkowski Spacetime: A Hundred Years Later* (Springer, Dordrecht 2010)

28.12 B.J. Carr: *Universe or Multiverse?* (Cambridge Univ. Press, Cambridge 2007)

28.13 P.S. Wesson: *Five-Dimensional Physics* (World Scientific, Singapore 2006)

28.14 P.S. Wesson: The geometrical unification of gravity with its source, J. Gen. Relativ. Gravit. **40**, 1353–1365 (2008)

28.15 P. Halpern: *The Great Beyond: Higher Dimensions, Parallel Universes, and the Extraordinary Search for a Theory of Everything* (Wiley, Hoboken 2004)

28.16 P. Halpern: *The Pursuit of Destiny: A History of Prediction* (Perseus, Cambridge 2000)

28.17 P. Halpern, P.S. Wesson: *Brave New Universe: Illuminating the Darkest Secrets of the Cosmos* (J. Henry, Washington 2006)

28.18 P.S. Wesson: *Weaving the Universe: Is Modern Cosmology Discovered Or Invented?* (World Scientific, Singapore 2011)

29. Viscous Universe Models

Øyvind Grøn, Diako Darian

We give a review of viscous relativistic universe models that have been presented during the period from 1990 until the present time. In particular we discuss the properties of isotropic and homogeneous universe models, and of anisotropic and homogeneous Bianchi type I models. We consider these types of models both in the context of the non-causal Eckhart theory and the causal Israel-Stewart theory.

The research on relativistic universe models with viscous fluids is reviewed. Viscosity may have been of significance during the early inflationary era, and may also be of importance for the late time evolution of the universe. Bulk viscosity and shear viscosity cause exponential decay of anisotropy, while nonlinear viscosity causes power-law decay of anisotropy. Redshift at transition from deceleration to acceleration is calculated.

29.1 Viscous Universe Models

Misner [29.1] noted that the

> *measurement of the isotropy of the cosmic background radiation represents the most accurate observational datum in cosmology,*

which is even more true today with the Wilkinson microwave anisotropy probe WMAP and Planck measurements. An explanation of this isotropy was provided by showing that in a large class of homogeneous but anisotropic universes, the anisotropy

dies away rapidly. It was found that the most important mechanism in reducing the anisotropy is neutrino viscosity at temperatures just above 10^{10} K (when the universe was about 1 s old: cf. *Zel'dovich* and *Novikov* [29.2]).

The first theory of relativistic viscous fluid was presented by *Eckart* in 1940 [29.3]. Eckart's theory deals with first-order deviation from equilibrium, while neglected second-order terms are necessary to prevent noncausal behavior. *Israel* and *Stewart* [29.4] have developed a second-order theory. *Grøn* [29.5] and *Maartens* [29.6, 7] have presented exhaustive reviews of research on cosmological models with noncausal and causal theories of viscous fluids, respectively.

Bulk viscosity-driven cosmic expansion with the Israel–Stewart theory have been investigated by *Zimdahl* [29.8], *Mak* and *Harko* [29.9], *Paul* et al. [29.10] and by *Arbab* and *Beesham* [29.11]. As noted by *Lepe* et al. [29.12], although Eckart's theory presents some causality problems, it is the simplest alternative and has been widely considered in cosmology, as documented in *Grøn*'s review [29.5], which we refer to for works on these topics up to 1990. We will here review papers from 1990 and onward.

Many types of observations favor that our universe is homogeneous and isotropic on scales above a billion light years. The observations of the temperature fluctuations in the cosmic microwave radiation favor that the universe is flat, i. e., that the total density of the matter and energy contained in the universe is equal to the critical density.

The discovery that the expansion of the universe accelerates could most simply be explained by repulsive gravity due to a cosmic vacuum energy with a density equal to about 70% of the critical density. The observations also favor a special type of vacuum energy, which may be represented by a cosmological constant in Einstein's field equations. The energy–momentum tensor of this energy is proportional to the metric tensor. One may show that this means that every component of the energy–momentum tensor is Lorentz invariant [29.13, 14]. Hence it is not possible to measure velocity with respect to this type of energy. It may therefore be called a Lorentz invariant vacuum energy, LIVE.

Furthermore, a large amount of cold dark matter is needed to keep the galaxies and the clusters of galaxies together because of the rapid motions of the stars in the galaxies and of the galaxies in the clusters. Hence, about 30% of the contents of the universe seem to be in the form of cold dark matter.

The cosmologists therefore introduced a standard model of the universe dominated by two fluids, a Lorentz invariant vacuum energy (LIVE), and a cold fluid. The vacuum energy is usually called dark energy and the cold fluid is called dark matter. Since the observations of the temperature fluctuations in the cosmic microwave radiation favor that the universe is flat, we shall only consider flat universe models.

29.2 The Standard Model of the Universe

For later comparison, we shall first briefly summarize the main properties of the standard model, which has vanishing viscosity. The total pressure and density are given by

$$\begin{aligned} \rho &= \rho_M + \rho_\Lambda , \\ p &= p_M + p_\Lambda = -\rho_\Lambda . \end{aligned} \tag{29.1}$$

The line-element has the form (using units so that the velocity of light in empty space is equal to 1),

$$\begin{aligned} \mathrm{d}s^2 &= -\mathrm{d}t^2 + a(t)^2(\mathrm{d}r^2 + r^2\,\mathrm{d}\Omega^2) , \\ \mathrm{d}\Omega^2 &= \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2 . \end{aligned} \tag{29.2}$$

The scale factor of this model is [29.15]

$$\begin{aligned} a(t) &= A^{1/3}\sinh^{2/3}\left(\frac{t}{t_\Lambda}\right) , \\ t_\Lambda &= 2/\sqrt{3\kappa\rho_\Lambda} , \quad A = \frac{1-\Omega_M}{\Omega_{\Lambda 0}} , \end{aligned} \tag{29.3}$$

where $\kappa = 8\pi G$ is Einstein's gravitational constant. Here $\Omega_{\Lambda 0}$ is the present value of the density parameter of LIVE. The scale factor represents the distance between two galaxy clusters relative to their present distance. Hence $a(t_0) = 1$ where the present age of the universe is

$$t_0 = t_\Lambda \operatorname{artanh}\sqrt{\Omega_{\Lambda 0}} . \tag{29.4}$$

Inserting the presently favored values $t_0 = 13.7\times10^9$ yr and $\Omega_{\Lambda 0} = 0.7$ leads to $t_\Lambda = 11.4\times10^9$ yr. The Hubble parameter is

$$H(t) = \frac{2}{3t_\Lambda}\coth\frac{t}{t_\Lambda}\,. \tag{29.5}$$

The deceleration parameter is

$$q(t) = \frac{1}{2}\left(1 - 3\tanh^2\frac{t}{t_\Lambda}\right). \tag{29.6}$$

The point of time t_I when deceleration turns into acceleration is given by $q(t_\mathrm{I}) = 0$ which leads to

$$t_\mathrm{I} = t_\Lambda \operatorname{artanh}\frac{1}{\sqrt{3}}\,. \tag{29.7}$$

The corresponding redshift is

$$z = \frac{1}{a(t_\mathrm{I})} - 1 = \left(\frac{2\Omega_{\Lambda 0}}{1-\Omega_{\Lambda 0}}\right)^{1/3} - 1\,, \tag{29.8}$$

which gives $t_\mathrm{I} = 7.4\times10^9$ yr and $z(t_\mathrm{I}) = 0.67$.

29.3 Viscous Fluid in an Expanding Universe

Let u^μ be components of the four-velocity of a fluid element. The projection tensor onto a 3-space orthogonal to the world line of a fluid element is defined by

$$h_{\alpha\beta} = g_{\alpha\beta} + u_\alpha u_\beta\,. \tag{29.9}$$

The covariant derivative of the velocity field of the fluid can be written as a 3×3 matrix which can be separated into the antisymmetric part representing the *vorticity*,

$$\omega_{\alpha\beta} = \frac{1}{2}\left(u_{\mu;\nu} - u_{\nu;\mu}\right) h^\mu_\alpha h^\nu_\beta\,, \tag{29.10}$$

the trace free, symmetrical part which represents the *shear*,

$$\sigma_{\alpha\beta} = \left[\frac{1}{2}\left(u_{\mu;\nu} - u_{\nu;\mu}\right) - \frac{1}{3}u^\lambda_{;\lambda} h_{\mu\nu}\right] h^\mu_\alpha h^\nu_\beta\,, \tag{29.11}$$

and the trace, which represents the *expansion*,

$$\theta = u^\lambda_{;\lambda}\,. \tag{29.12}$$

The four acceleration of the fluid element, which is nonvanishing only for nongeodesic flow, is defined by

$$a_\alpha = a_{\alpha;\mu} u^\mu\,. \tag{29.13}$$

We then have

$$u_{\alpha;\beta} = \omega_{\alpha\beta} + \sigma_{\alpha\beta} + \tfrac{1}{3}\theta h_{\mu\nu} - a_\alpha u_\beta\,. \tag{29.14}$$

The energy–momentum tensor of a viscous fluid with proper density ρ and pressure p is

$$T_{\alpha\beta} = \rho u_\alpha u_\beta + (p - \xi\theta)\, h_{\alpha\beta} - 2\eta\sigma_{\alpha\beta}\,, \tag{29.15}$$

where η and ξ are the coefficients of shear and bulk viscosity, respectively. Einstein's field equations imply that the divergence of this tensor vanishes. From this one may deduce the equation of continuity of the fluid in the form [29.16]

$$\dot\rho + (\rho + p)\,\theta - 4\eta\sigma^2 - \xi\theta^2 = 0\,. \tag{29.16}$$

The evolution of the divergence with time is given by the Raychaudhuri equation which may be written as [29.16]

$$\dot\theta = a^\lambda_{;\lambda} + \frac{\kappa}{2}\left(3\xi\theta - \rho - 3p\right) - 2\omega^2 - 2\sigma^2 - \frac{1}{3}\theta^2\,. \tag{29.17}$$

29.4 Isotropic, Viscous Generalization of the Standard Universe Model

We now consider a homogeneous and isotropic universe with geodesic fluid flow. In this case $a^{\lambda}_{;\lambda} = \omega = \sigma = 0$ and (29.16) and (29.17) reduce to

$$\dot{\rho} + (\rho + p)\theta - \xi\theta^2 = 0\,, \tag{29.18}$$

and

$$\dot{\theta} = \frac{\kappa}{2}(3\xi\theta - \rho - 3p) - \frac{1}{3}\theta^2\,. \tag{29.19}$$

Assuming that the 3-space is Euclidean, the line element has the form given by (29.2). As a further generalization of the standard model we shall assume that the universe contains two noninteracting fluids, Lorentz invariant vacuum energy, LIVE, that may be represented by a cosmological constant, $\kappa\rho_\Lambda = -\kappa p_\Lambda = \Lambda$, and a fluid with equation of state $p_m = w\rho_m$, so that $\rho = \rho_m + \rho_\Lambda$. *Ren* and *Meng* [29.17] and *Hu* and *Meng* [29.18] have proposed the following form of the bulk viscosity coefficient:

$$\xi = \xi_0 + \xi_1\frac{\dot{a}}{a} + \xi_2\frac{\ddot{a}}{\dot{a}}\,. \tag{29.20}$$

The motivation for considering this form for the coefficient of bulk viscosity is that from fluid mechanics we know that the viscosity is related to the motion of the fluid, i. e., to $\dot{a}$ and $\ddot{a}$.

With the metric (29.2) the expansion is $\theta = 3\dot{a}/a = 3H$, where $H \equiv \dot{a}/a$ is the Hubble parameter. Inserting (29.20) into (29.18) the equation of continuity takes the form

$$\dot{\rho} + 3(1+w)\rho H - 9\xi_0 H^2 + 9\xi_1 H^3 + 9\xi_2(H\dot{H} + H^3)H^2 = 0\,. \tag{29.21}$$

Einstein's field equations take the form

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{\kappa}{3}\rho\,, \tag{29.22}$$

$$\frac{\ddot{a}}{a} = -\frac{\kappa}{6}(\rho + 3p - 9\xi H)\,. \tag{29.23}$$

Inserting (29.22) into (29.19), the Raychaudhuri equation takes the form

$$\dot{H} = -\frac{3}{2}H^2 + \frac{3}{2}\kappa\xi H + \frac{\kappa}{2}\rho_\Lambda\,. \tag{29.24}$$

We define the density parameters

$$\Omega_{\Lambda 0} = \frac{\kappa\rho_\Lambda}{3H_0^2}\,,\quad \Omega_{m0} = \frac{\kappa\rho_{m0}}{3H_0^2}\quad\text{and}\quad \Omega_{\xi 0} = \frac{3\kappa\xi_0}{H_0}\,, \tag{29.25}$$

where $H_0 = [(\kappa/3)\,\rho_0]^{1/2}$.

The field equations can be integrated analytically for some special cases [29.19]. Choosing $\xi_1 = \xi_2 = 0$ and integrating (29.24) two times with the boundary conditions $a(0) = 0$ and $a(t_0) = 1$ we obtain

$$H(t) = \frac{\kappa\xi_0}{2} + \hat{H}\coth\left(\frac{3}{2}\hat{H}t\right) \tag{29.26}$$

with

$$\hat{H} = H_0\sqrt{\left(\frac{\Omega_{\xi 0}}{2}\right)^2 + \Omega_{\Lambda 0}} \tag{29.27}$$

and

$$a(t) = A e^{\frac{\kappa\xi_0}{2}(t-t_0)}\sinh^{2/3}\left(\frac{3}{2}\hat{H}t\right)\,,\quad A = \left(\frac{\Omega_{m0} - \Omega_{\xi 0}}{\Omega_{\Lambda 0} + \left(\frac{\Omega_{\xi 0}}{2}\right)^2}\right)^{1/3}\,. \tag{29.28}$$

From (29.26) it follows that the age of this universe model is

$$t_0 = \frac{2}{3\hat{H}}\operatorname{artanh}\frac{\sqrt{\left(\frac{\Omega_{\xi 0}}{2}\right)^2 + \Omega_{\Lambda 0}}}{1 - \frac{\Omega_{\xi 0}}{2}}\,. \tag{29.29}$$

The corresponding age of the universe if the viscosity vanishes is given by (29.4). Hence for a given present value of the Hubble parameter the viscosity increases the age of the universe. Assuming that $\kappa\xi_0 \ll H_0$, the increase of the age due to the viscosity is approximately

$$t_0 - t_{00} \approx \frac{1}{3H_0}\frac{\left(\frac{\Omega_{\xi 0}}{2}\right)^2}{\Omega_{\Lambda 0}\,(1 - \Omega_{\Lambda 0})}\,. \tag{29.30}$$

Brevik and *Heen* [29.20] have used that in the plasma era of the universe the bulk viscosity derived from kinetic theory of gases has order of magnitude so that $(\kappa\xi_0)^{-1} \approx 10 \times 10^{21}$ yr. Since $(H_0)^{-1} \approx 10^{10}$ years this estimate of the magnitude of the bulk viscosity gives $\Omega_{\xi 0} \approx 10^{-11}$. During most of the evolution of the universe the viscosity is smaller than this. Hence, this form of viscosity is totally insignificant for the age of the universe.

Brevik and *Heen* [29.20] have, however, pointed out that impulsive processes at the end of the inflationary era may have produced great viscosity. In this extremely brief period the viscosity may have given significant contributions to the production of entropy in the universe, able to explain why the number of photons per baryon is so large, $\approx 10^9$, in our universe.

The continuity equation for matter takes the form

$$\dot{\rho_m} = -3H\left(\rho_m - 3\xi_0 H\right). \qquad (29.31)$$

Inserting (29.26) into (29.31) and integrating both sides of this equation, yields

$$\rho_m(t) = 3\xi_0 H(t) + \left[\rho_{m0} - 3\xi_0 H(t)\right] \mathrm{e}^{-\frac{3}{2}\kappa\xi_0(t-t_0)} \times \frac{1}{A^3 \sinh^2\left(\frac{3}{2}\hat{H}t\right)}, \qquad (29.32)$$

where ρ_{m0} is the energy density of matter at the present time, t_0. Next we insert (29.20) into (29.24) to obtain

$$a\dot{H} = bH^2 + cH + d, \qquad (29.33)$$

where

$$\begin{aligned} a &= 1 - \frac{3}{2}\kappa\xi_2, \\ b &= \frac{3}{2}\left(\kappa(\xi_1 + \xi_2) - 1\right), \\ c &= \frac{3}{2}\kappa\xi_0, \\ d &= \frac{1}{2}\kappa\rho_\Lambda. \end{aligned} \qquad (29.34)$$

Integration with $a(0) = 0$ and $a(t_0) = 1$ gives

$$H(t) = -\frac{c}{b} + \hat{H}\coth\left(-\frac{b}{a}\hat{H}t\right) \qquad (29.35)$$

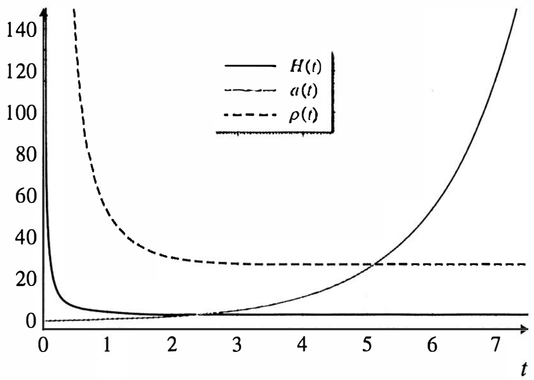


Fig. 29.1 The dynamical evolution of the Hubble parameter, the scale factor, and the energy density as functions of time. Here $\kappa\xi_0 = 0.1$, $\kappa\xi_1 = 0.5$ and $\kappa\xi_2 = 0.4$

and

$$a(t) = \mathrm{e}^{-\frac{c}{2b}(t-t_0)}\left(\frac{\sinh\left(-\frac{b}{a}\hat{H}t\right)}{\sinh\left(-\frac{b}{a}\hat{H}t_0\right)}\right)^{-b/a}, \qquad (29.36)$$

where

$$\hat{H}^2 = \left(\frac{c}{2b}\right)^2 - \frac{\mathrm{d}}{b}. \qquad (29.37)$$

In Fig. 29.1, we have plotted the evolution of the Hubble parameter, the scale factor, and the energy density as functions of time for some values of the bulk viscosity coefficients. This figure shows that the scale factor starts with zero and the Hubble parameter and the energy density are infinitely large at beginning of the cosmic evolution, which shows that there is a singularity at the initial epoch, and therefore the universe starts with a Big Bang. As t increases the scale factor will increase exponentially, and, as $t \to \infty$, it becomes infinite, whereas the Hubble parameter and the energy density become finite. If the bulk viscosity is zero, the energy density tends to zero as $t \to \infty$. Therefore, this model will give an empty universe for large times. For the ΛCDM model with bulk viscosity, as $t \to \infty$ the energy density converges to a finite value. It means that for this model the energy density will stay constant for large times. The bigger the value of the bulk viscosity coefficient ξ_0 is, the bigger this constant value of the energy density will be.

29.5 The Dark Sector of the Universe as a Viscous Fluid

In the standard ΛCDM model, the universe today is dynamically dominated by a dark sector which consists of dark matter and dark energy, and occupies 96% of its total energy content. The dark energy in the ΛCDM model is represented by a cosmological constant, Λ. One problem with the cosmological constant is that its theoretical value is between 60–120 orders of magnitude greater than the observed value. Another problem in the ΛCDM model is the dark degeneracy problem [29.21, 22], which is the disability of the present gravitational probes to differentiate dark matter from dark energy. Therefore, it is reasonable to model dark matter and dark energy with a single fluid.

There are some models of this type, like the unified dark energy model and models with ordinary and generalized Chaplygin gas (see [29.23] and references therein). Here, we will concentrate on a description of the dark sector with a viscous fluid. We will, therefore, look at two different models for a unified dark fluid with viscosity.

29.5.1 Bulk Viscosity as a Model for Unified Dark Matter with the EoS $p = (\gamma - 1)\rho$

We consider the Friedmann–Robertson–Walker universe with metric given in (29.2) and energy–momentum tensor given by (29.15). The Friedmann equations and the continuity equation are

$$3H^2 = \kappa\rho, \tag{29.38}$$

$$\frac{\ddot{a}}{a} = -\frac{\kappa}{6}\left(\rho + 3(p + \Pi)\right), \tag{29.39}$$

$$\dot{\rho} + 3\left(\rho + p + \Pi\right)H = 0\,. \tag{29.40}$$

Here we will use the Eckart theory and set $\Pi = -3\xi H$. Inserting the EoS into (29.40), we obtain

$$\dot{\rho} + 3\left(\gamma\rho - 3\xi H\right)H = 0\,. \tag{29.41}$$

Defining the dimensionless Hubble parameter as

$$h \equiv \frac{H}{H_0}\,, \tag{29.42}$$

where H_0 is the present value of the Hubble parameter, (29.38) and (29.41) can be rewritten as

$$h^2 = \frac{\rho}{\rho_{\mathrm{cr}}}\,, \tag{29.43}$$

$$\frac{1}{H_0}\frac{\mathrm{d}(h^2)}{\mathrm{d}t} + 3\gamma h^3 = 9\lambda h^2\,, \tag{29.44}$$

where $\rho_{\mathrm{cr}} = \frac{3H_0^2}{\kappa}$ is the critical density and $\lambda \equiv \frac{H_0\xi}{\rho_{\mathrm{cr}}}$. Using the transformation

$$\mathrm{d}t = \frac{1}{aH}\mathrm{d}a\,, \tag{29.45}$$

we rewrite (29.44) as

$$\frac{\mathrm{d}H}{\mathrm{d}a} + \frac{3\gamma}{2a}H = \frac{3\kappa\xi}{2a}\,. \tag{29.46}$$

Integrating (29.46), we obtain

$$H(a) = Ca^{-\frac{3\gamma}{2}} + \left[\int \frac{3\kappa\xi}{2a}\exp\left(\int \frac{3\gamma}{2a}\mathrm{d}a\right)\mathrm{d}a\right] \times \exp\left(-\int \frac{3\gamma}{2a}\mathrm{d}a\right). \tag{29.47}$$

Depending on the form of the bulk viscosity, ξ, this integral can be solved analytically or numerically. Combining (29.38) and (29.40), we obtain the Raychaudhuri equation

$$\dot{H} = -\frac{3}{2}\gamma\left(H - \frac{\kappa\xi}{\gamma}\right)H\,. \tag{29.48}$$

Integration of this equation with the assumption that the bulk viscosity is constant, gives

$$H(t) = \frac{\kappa\xi}{\gamma}\frac{1}{1 - \left(1 - \frac{\kappa\xi}{\gamma H_0}\right)\mathrm{e}^{-\frac{3}{2}\kappa\xi(t-t_0)}}\,. \tag{29.49}$$

Using $H = \frac{\dot{a}}{a}$, and integrating (29.49) with $a(0) = 0$ and $a(t_0) = 1$, we get the following expression for the scale factor:

$$a(t) = \mathrm{e}^{\frac{\kappa\xi}{\gamma}(t-t_0)}\left[\frac{\gamma H_0}{\kappa\xi}\left(1 - \mathrm{e}^{-\frac{3}{2}\kappa\xi t}\right)\right]^{\frac{2}{3\gamma}}\,, \tag{29.50}$$

where

$$t_0 = -\frac{2}{3\kappa\xi}\ln\left(1 - \frac{\kappa\xi}{\gamma H_0}\right). \tag{29.51}$$

The energy density as a function of time is then

$$\rho(t) = \frac{3\kappa\xi^2}{\gamma^2}\frac{1}{\left(1-e^{-\frac{3}{2}\kappa\xi t}\right)^2} . \qquad (29.52)$$

These equations describe a universe that begins in a Big Bang and evolves to have constant energy density and expands forever with a constant Hubble parameter. In what follows we will look at two universe models with different bulk viscosity parameter ξ [29.23–25].

Model I: $\xi = \xi_0 + \xi_1\frac{\dot{a}}{a} + \xi_2\frac{\ddot{a}}{\dot{a}}$

A universe model with this bulk viscosity parameter is equivalent with a model that has the effective equation of state

$$p = (\gamma-1)\rho + p_0 + w_H H + w_{H_2}H^2 + w_{dH}\dot{H} , \qquad (29.53)$$

where p_0, w_H, w_{H_2}, and w_{dH} are free parameters. The equivalence between these two models is given by the following transformation:

$$w_H = -3\xi_0 , \qquad (29.54a)$$

$$w_{H_2} = -3(\xi_1+\xi_2) , \qquad (29.54b)$$

$$w_{dH} = -3\xi_2 . \qquad (29.54c)$$

Combining (29.38) and (29.39), we get the following differential equation for the Hubble parameter:

$$\dot{H} = k_1H^2 + k_2H , \qquad (29.55)$$

where

$$k_1 = \frac{\frac{3}{2}\left[\kappa(\xi_1+\xi_2)-\gamma\right]}{1-\frac{3}{2}\kappa\xi_2} ,$$

$$k_2 = \frac{\frac{3}{2}\kappa\xi_0}{1-\frac{3}{2}\kappa\xi_2} .$$

Integration of (29.55) gives

$$H(t) = \frac{k_2H_0e^{k_2(t-t_0)}}{k_2+k_1H_0\left(1-e^{k_2(t-t_0)}\right)} . \qquad (29.56)$$

The scale factor takes the form

$$a(t) = e^{-\frac{k_2}{k_1}(t-t_0)} \times \left(\frac{k_2}{(k_1H_0+k_2)e^{-k_2(t-t_0)}-k_1H_0}\right)^{1/k_1} . \qquad (29.57)$$

From Friedmann's first equation, i. e., (29.38), the energy density can be written as

$$\rho(t) = \frac{3}{\kappa}\left(\frac{k_2H_0e^{k_2(t-t_0)}}{k_2+k_1H_0\left(1-e^{k_2(t-t_0)}\right)}\right)^2 . \qquad (29.58)$$

The deceleration parameter is defined as

$$q = -\frac{\ddot{a}}{aH^2} = -1-\frac{\dot{H}}{H^2} . \qquad (29.59)$$

By using (29.56), q takes the form

$$q(t) = -1-\frac{(k_1H_0+k_2)e^{-k_2(t-t_0)}}{H_0} . \qquad (29.60)$$

Rewriting the bulk viscosity as

$$\begin{aligned}\xi &= \xi_0+\xi_1\frac{\dot{a}}{a}+\xi_2\frac{\ddot{a}}{\dot{a}}\\ &= \xi_0+(\xi_1-\xi_2q)H ,\end{aligned}$$

we obtain the evolution of bulk viscosity

$$\xi(t) = \frac{(k_2+k_1H_0)(\xi_0+k_2\xi_2)+\left[k_1\xi_0+k_2(\xi_1+\xi_2)\right]H_0e^{k_2(t-t_0)}}{k_2+k_1H_0\left(1-e^{k_2(t-t_0)}\right)} . \qquad (29.61)$$

Model II: $\xi = \xi_0 + \xi_1 H + \xi_2 H^2$

In this model the Raychaudhuri equation takes the form

$$\dot{H} = \frac{3}{2}\kappa\xi_2H\left[(H+b)^2-\hat{H}^2\right] , \qquad (29.62)$$

where

$$\hat{H}^2 \equiv b^2-\frac{\xi_0}{\xi_2} \quad \text{and} \quad b \equiv \frac{\kappa\xi_1-\gamma}{2\kappa\xi_2} . \qquad (29.63)$$

Integration of (29.62) gives

$$t(H) = t_0-\frac{2}{3\kappa\xi_2}\ln\left[\left(\frac{H_0}{H}\right)^{\frac{\xi_2}{\xi_0}}\left(\frac{H_0+b-\hat{H}}{H+b-\hat{H}}\right)^{\frac{1}{2\hat{H}(\hat{H}-b)}} \times\left(\frac{H_0+b+\hat{H}}{H+b+\hat{H}}\right)^{\frac{1}{2\hat{H}(\hat{H}+b)}}\right] . \qquad (29.64)$$

Equation (29.64) is very complicated and does not give an expression for the Hubble parameter. Therefore, we define $x \equiv \ln a$, and we use the transformation

$$\frac{d}{dt} = H\frac{d}{dx}\,, \tag{29.65}$$

to rewrite (29.44) as

$$h' + \frac{3}{2}\gamma h = \frac{3}{2}\frac{\kappa}{H_0^2}\xi\,. \tag{29.66}$$

Inserting the expression for the bulk viscosity into this equation, we obtain

$$h' = \frac{3}{2}\kappa\xi_2 H_0 h^2 + \frac{3}{2}(\kappa\xi_1 - \gamma)h + \frac{3}{2}\frac{\kappa\xi_0}{H_0}\,, \tag{29.67}$$

This equation can be integrated, and the solution depends on the sign of

$$\Delta \equiv \frac{9}{16}(\kappa\xi_1 - \gamma)^2 - \frac{9}{4}\kappa^2\xi_0\xi_2\,. \tag{29.68}$$

The solutions are

- $\Delta < 0$:

$$h(x) = -\frac{c_2}{c_1} + \frac{\sqrt{-\Delta}}{c_1}\tan\left(\phi + \sqrt{-\Delta}x\right), \tag{29.69}$$

  where $\phi \equiv \arctan\frac{c_1+c_2}{\sqrt{-\Delta}}$, $c_1 \equiv \frac{3}{2}\kappa\xi_2 H_0$ and $c_2 \equiv \frac{3}{2}(\kappa\xi_1 - \gamma)$.
- $\Delta = 0$:

$$h(x) = \frac{1}{c_1}\left(\frac{1}{x} - c_2\right). \tag{29.70}$$

- $\Delta > 0$:

$$h(x) = \frac{1}{c_1}\frac{(c_2 + \sqrt{\Delta})Ke^{2\sqrt{\Delta}x} - c_2 + \sqrt{\Delta}}{1 - Ke^{2\sqrt{\Delta}x}}\,, \tag{29.71}$$

  where $K \equiv \frac{c_1+c_2-\sqrt{\Delta}}{c_1+c_2+\sqrt{\Delta}}$.

Meng and *Ma* [29.25] have constrained these models with the latest Union2 data [29.26] and the currently observed Hubble-parameter dataset (OHD) [29.27]. From their results by best fitting they have found that these models are consistent with the observational data in the region of data fitting, but because of the presence of bulk viscosity these models have a much more flexible evolution processing. They conclude that with the bulk viscosity considered, a more realistic universe scenario is obtained comparable with the ΛCDM model but without introducing the mysterious dark energy.

Furthermore, they have found that Model I mimics the evolution of the ΛCDM universe model perfectly. Since Model I shares many similarities with the ΛCDM model, it makes this kind of bulk viscosity parameter a successful substitution for the dark energy model. For Model II they have only studied the solutions $\Delta > 0$ and $\Delta < 0$. For the case $\Delta < 0$ the universe is bounded, and the expansion ends at far future $t = t_c$, when H and $\dot{H}$ both vanishes. However, this solution fits the data quite well in the corresponding data region. For the case $\Delta > 0$ the universe does not have a Big Bang scenario. But, in the region of data fitting this universe model mimics the observed acceleration successfully.

29.5.2 Unified Dark Matter with the EoS $p = -\xi\theta$

In this section, we will look at a unified description of dark matter and dark energy proposed by *Hipolito-Ricaldi* et al. [29.28]. We assume that the dark sector of the cosmic substratum is a viscous fluid with the equation of state

$$p = -\xi\theta\,, \tag{29.72}$$

where we assume that the bulk viscosity coefficient is given by

$$\xi = \xi_0\rho^\nu\,, \tag{29.73}$$

where ξ_0 and ν are positive constants. Inserting (29.38) and (29.73) into (29.72) and using that $\theta = 3H$, we obtain

$$p = -\xi_0\theta\rho^\nu = -A\rho^{\nu+1/2}\,, \tag{29.74}$$

where $A = \sqrt{3\kappa}\xi_0 > 0$. This equation can be compared to the equation of state of a generalized Chaplygin gas, which has the form

$$p_{gcg} = -\frac{A}{\rho^\alpha}\,, \tag{29.75}$$

and we obtain the correspondence $\alpha = -(\nu + 1/2)$. For the study of the similarity between generalized Chaplygin gases and bulk viscous fluids see [29.29, 30]. For $\nu = 1/2$ and $\alpha = -1$ and $A = 1$ both models contain the ΛCDM model as a special case.

Friedmann's first equation takes the form

$$\theta^2 = 3\kappa\rho\,. \tag{29.76}$$

With the equation of state

$$\frac{p}{\rho} = \gamma - 1\,, \tag{29.77}$$

where γ is not constant, (29.16) and (29.17) reduce to

$$\dot\rho + \theta\,(\rho + p) = 0\,, \tag{29.78}$$

and

$$\dot\theta = -\frac{\kappa}{2}\,(\rho + 3p) - \frac{1}{3}\theta^2 \quad \Longrightarrow \quad \dot\theta = -\frac{\gamma}{2}\theta^2\,. \tag{29.79}$$

Integrating (29.78), we obtain the following relationship between the energy density and the scale factor:

$$\rho = \left[A + B\left(\frac{a_0}{a}\right)^{\frac{3}{2}(1-2\nu)}\right]^{\frac{2}{1-2\nu}}\,, \tag{29.80}$$

where $B \equiv \frac{1 - A\rho_0^{\nu+1/2}}{\rho_0^{\nu+1/2}}$ and a_0 and ρ_0 are the present values of the scale factor and the energy density, respectively. Inserting (29.80) into (29.76), gives

$$H = \sqrt{\frac{\kappa}{3}}\left[A + B\left(\frac{a_0}{a}\right)^{\frac{3}{2}(1-2\nu)}\right]^{\frac{1}{1-2\nu}}\,. \tag{29.81}$$

Using the equation $q = -1 - \frac{\dot H}{H^2}$, we obtain the following equation for the deceleration parameter:

$$q = -\frac{1 - \frac{B}{2A}\left(\frac{a_0}{a}\right)^{\frac{3}{2}(1-2\nu)}}{1 + \frac{B}{A}\left(\frac{a_0}{a}\right)^{\frac{3}{2}(1-2\nu)}}\,. \tag{29.82}$$

The present value of the deceleration parameter is

$$q_0 = -\frac{1 - \frac{B}{2A}}{1 + \frac{B}{A}}\,, \tag{29.83}$$

which we rewrite as

$$\frac{B}{2A} = \frac{1 + q_0}{1 - 2q_0}\,. \tag{29.84}$$

At $q = 0$ we have the transition from decelerated to accelerated expansion. Using (29.82), we find the following value for the scale factor at this point, which we denote by a_{acc}:

$$a_{\mathrm{acc}} = a_0\left(\frac{B}{2A}\right)^{\frac{2}{3(1-2\nu)}}\,. \tag{29.85}$$

By using the relation $1 + z = \frac{a_0}{a}$, we obtain the corresponding redshift

$$z_{\mathrm{acc}} = \left(\frac{1 - 2q_0}{1 + q_0}\right)^{\frac{2}{3(1-2\nu)}} - 1\,. \tag{29.86}$$

Expressing the Hubble parameter and the energy density as functions of q_0, we obtain

$$\frac{H}{H_0} = \left(\frac{1}{3}\right)^{\frac{1}{1-2\nu}} \times\left[1 - 2q_0 + 2\,(1 + q_0)\left(\frac{a_0}{a}\right)^{\frac{3}{2}(1-2\nu)}\right]^{\frac{1}{1-2\nu}}\,, \tag{29.87}$$

and

$$\frac{\rho}{\rho_0} = \left(\frac{1}{9}\right)^{\frac{2}{1-2\nu}} \times\left[1 - 2q_0 + 2\,(1 + q_0)\left(\frac{a_0}{a}\right)^{\frac{3}{2}(1-2\nu)}\right]^{\frac{2}{1-2\nu}}\,. \tag{29.88}$$

The equation of state parameter is then given by

$$\gamma = 1 + \frac{p}{\rho} = \frac{2\,(1 + q_0)\left(\frac{a_0}{a}\right)^{\frac{3}{2}(1-2\nu)}}{1 - 2q_0 + 2\,(1 + q_0)\left(\frac{a_0}{a}\right)^{\frac{3}{2}(1-2\nu)}}\,. \tag{29.89}$$

Hipolito-Ricaldi et al. [29.28] have calculated the matter power spectrum for this model and they have concluded that unified models with bulk viscosity with $\xi \propto \rho^{\nu}$ are compatible with the current observational data. Also, for certain parameter combinations, their χ^2 analysis favors the unified viscous dark fluid model over the standard ΛCDM universe model.

29.6 Viscosity and the Accelerated Expansion of the Universe

The question whether a matter dominated universe with the constant bulk viscosity can drive the accelerated expansion of the universe has been discussed by *Avelino* and *Nucamendi* [29.31]. In this universe model $\Omega_m = 1$, $\Omega_\Lambda = 0$, and the expression (29.28) for the scale factor reduces to

$$a(t) = \left(\frac{4\left(1-\Omega_{\xi 0}\right)}{\Omega_{\xi 0}^2}\right)^{1/3} e^{(\Omega_{\xi 0}/2)H_0(t-t_0)} \times \sinh^{2/3}\left(\frac{3}{4}\Omega_{\xi 0}H_0 t\right) . \tag{29.90}$$

where the age of the universe is

$$\begin{aligned} t_0 &= \frac{4}{3\Omega_{\xi 0}H_0}\operatorname{artanh}\frac{\Omega_{\xi 0}}{2-\Omega_{\xi 0}} \\ &= -\frac{2}{3\Omega_{\xi 0}H_0}\ln\left(1-\Omega_{\xi 0}\right) . \end{aligned} \tag{29.91}$$

This form of the solution satisfies the boundary conditions $a(0) = 0$, $a(t_0) = 1$. The first of these are not satisfied by the form of the solution given by *Avelino* and *Nucamendi* [29.31]. Solution (29.90) may be written as

$$a(t) = \left(\frac{1-\Omega_{\xi 0}}{\Omega_{\xi 0}}\right)^{2/3} \left(e^{\frac{3}{2}\Omega_{\xi 0}H_0 t} - 1\right)^{2/3} . \tag{29.92}$$

The Hubble parameter is

$$H(t) = \frac{\Omega_{\xi 0}H_0}{1 - e^{-\frac{3}{2}\Omega_{\xi 0}H_0 t}} . \tag{29.93}$$

This universe model approaches the de Sitter model for $t \gg 1/\Omega_{\xi 0}H_0$ with a constant Hubble parameter equal to $\Omega_{\xi 0}H_0$. The deceleration parameter is in general given by

$$q = -1 - \frac{\dot{H}}{H^2} . \tag{29.94}$$

The deceleration parameter of the universe models considered here is

$$q(t) = \frac{3}{2e^{\frac{3}{2}\Omega_{\xi 0}H_0 t}} - 1 . \tag{29.95}$$

The present value of the deceleration parameter is

$$q(t_0) = \tfrac{1}{2}\left(1 - 3\Omega_{\xi 0}\right) . \tag{29.96}$$

These expressions show that the expansion starts from a Big Bang with an infinitely great expansion velocity, but decelerates to a finite value. At a point of time t_1 given by $q(t_1) = 0$ there is a transition to accelerated expansion, which will last forever. The transition happens at

$$t_1 = \frac{2\ln\frac{3}{2}}{3\Omega_{\xi 0}H_0} . \tag{29.97}$$

At this point of time the scale factor has the value

$$a(t_1) = \left(\frac{1-\Omega_{\xi 0}}{2\Omega_{\xi 0}}\right)^{2/3} . \tag{29.98}$$

The corresponding redshift is

$$z_1 = \left(\frac{2\Omega_{\xi 0}}{1-\Omega_{\xi 0}}\right)^{2/3} - 1 . \tag{29.99}$$

In order that the transition shall have happened at a past time, $a(t_1) < 1$, the bulk viscosity must be sufficiently great, $\Omega_{\xi 0} > 1/3$.

For this universe model, with Euclidean spatial geometry, the matter density is equal to the critical density,

$$\rho_m = \frac{3H^2}{\kappa} = \frac{3\Omega_{\xi 0}^2 H_0^2}{\kappa\left(1 - e^{-\frac{3}{2}\Omega_{\xi 0}H_0 t}\right)^2} . \tag{29.100}$$

Hence, the matter density approaches a constant value $\rho_m \to (3/\kappa)\Omega_{\xi 0}^2 H_0^2$.

Avelino and *Nucamendi* [29.31] have used the most comprehensive supernova data to estimate the value of $\Omega_{\xi 0}$ that gives the best fit with observed data for a universe model containing dust with constant coefficient of viscosity. The result was $\Omega_{\xi 0} = 0.64$, which is 11 orders of magnitude greater than the value coming from kinetic gas theory [29.20]. However a mechanism for producing greater viscosity may be generation of bulk viscosity due to decay of dark matter particles into relativistic products [29.32, 33].

29.7 Viscous Universe Models with Variable G and Λ

There have been proposed many universe models in which the gravitational parameter G varies with the cosmic time, see [29.34–42] and references therein. Since G couples geometry to matter, it is reasonable to expect that in an evolving universe we might have $G = G(t)$. In this section we will briefly review the spatially flat Friedmann–Robertson–Walker universe models containing viscous fluid with variable gravitational coupling constant and variable cosmological constant. The Friedmann–Robertson–Walker space-time is given by the metric in (29.2). The Friedmann equations take the form

$$3H^2 = 8\pi G\rho + \Lambda\,, \tag{29.101}$$

$$3\frac{\ddot{a}}{a} = -4\pi G\left[\rho + 3\left(p + \Pi\right)\right] + \Lambda\,, \tag{29.102}$$

where Π is the bulk viscous pressure. The deceleration parameter is given by

$$q = -\frac{\ddot{a}}{aH^2} = -1 - \frac{\dot{H}}{H^2}\,. \tag{29.103}$$

A universe with accelerating expansion has negative deceleration parameter. From the Bianchi identities we obtain

$$\dot{\rho} + 3\left(p + \rho + \Pi\right)H = -\frac{\dot{G}}{G}\rho - \frac{\dot{\Lambda}}{8\pi G}\,. \tag{29.104}$$

Assuming there is no creation of particles, we can rewrite (29.104) as

$$\dot{\rho} + 3\left(p + \rho + \Pi\right)H = 0\,, \tag{29.105}$$

and

$$8\pi\dot{G}\rho + \dot{\Lambda} = 0\,. \tag{29.106}$$

Using the equation of state $p = w\rho$ the Raychaudhuri equation reduces to

$$\dot{H} = -\frac{3}{2}(1+w)H^2 - 4\pi G\Pi + \frac{(7+3w)}{6}\Lambda\,. \tag{29.107}$$

In order to solve these equations we assume $\Lambda = 3mH^2$, where m is a constant. In what follows, we explore universe models with a power law expansion given by

$$a(t) = a_0 t^{\nu}\,, \tag{29.108}$$

where a_0 and ν are constants. For a universe with accelerating expansion (i. e., $q < 0$) it follows from (29.103) that $\nu > 1$. From (29.108), we obtain directly

$$H(t) = \frac{\nu}{t}\,, \tag{29.109}$$

and

$$\Lambda(t) = \frac{3m\nu^2}{t^2}\,. \tag{29.110}$$

Combining (29.101) and (29.106), we obtain the following differential equation for G:

$$\frac{\dot{G}}{G} = \frac{2m}{1-m}\frac{1}{t}\,. \tag{29.111}$$

Integrating (29.111), we get

$$G(t) = G_0\left(\frac{t}{t_0}\right)^{\frac{2m}{1-m}}\,, \quad m \neq 1\,, \tag{29.112}$$

where G_0 is the value of the gravitational parameter at the present time t_0. From this equation, we see that for $0 < m < 1$ the gravitational parameter increases with time and Λ decreases, and it decreases for $m > 1$. It is evident that G is a constant when Λ vanishes. The energy density can now be obtained from (29.101)

$$\rho(t) = \frac{3(1-m)\nu^2 t_0^{\frac{2m}{1-m}}}{8\pi G_0}\frac{1}{t^{2\frac{1+m}{1-m}}}\,. \tag{29.113}$$

Assuming that $\rho > 0$ gives the constraint $m < 1$ which sets the upper boundary for the cosmological parameter to $\Lambda < 3H^2$. From the continuity equation, we obtain the following expression for the bulk viscous pressure:

$$\Pi(t) = \frac{\nu\left[2(1+m) - 3(1-m)(1+w)\nu\right] t_0^{\frac{2m}{1-m}}}{8\pi G_0} \times \frac{1}{t^{2\frac{1+m}{1-m}}}\,. \tag{29.114}$$

The bulk viscous pressure is assumed to be negative, i. e., $\Pi < 0$, which demands $\nu > \frac{2(1+m)}{3(1-m)(1+w)}$. For $-1 < m < 1$ the bulk viscous pressure will decrease with time. In the Eckart theory the bulk viscosity pressure is given by

$$\Pi = -3\xi H\,, \tag{29.115}$$

where ξ is the bulk viscosity parameter. From (29.109) and (29.114), we obtain

$$\xi = \xi_0 \rho^{\alpha}\,, \tag{29.116}$$

where

$$\xi_0 = -\frac{2(1+m)-3(1-m)(1+w)\nu}{9\nu^2(1-m)} \times \left(\frac{3(1-m)\nu^2 t_0^{\frac{2m}{1-m}}}{8\pi G_0}\right)^{\frac{1-m}{2(1+m)}}\,, \tag{29.117}$$

and

$$\alpha = \frac{1+3m}{2(1+m)}\,. \tag{29.118}$$

For a positive cosmological parameter and with $m < 1$ we have the constraint $1/2 < \alpha < 1$.

29.8 Hubble Parameter in the QCD Era of the Early Universe in the Presence of Bulk Viscosity

Tawfik et al. [29.43] have studied the evolution of the Hubble parameter in the QCD era of the early universe in the presence of viscous QCD plasma. Based on the recent lattice QCD simulations and heavy-ion collisions [29.44, 45], they have approximately determined the equation of state, the temperature and the bulk viscosity of the quark-gluon plasma and used this information to calculate the evolution of the universe in this era. They have used both the Eckart and the Israel–Stewart theory. The bulk viscosity is given by

$$\xi = \xi_0 \rho + b\,, \tag{29.119}$$

where ξ_0 and b are constants. The universe is assumed to be homogeneous and isotropic, and it contains matter that is given by the barotropic equation of state, obtained from lattice QCD simulations and heavy-ion collisions [29.44, 45]. The Friedmann equations and the continuity equation take the form

$$3H^2 = \kappa\rho\,, \quad 3\frac{\ddot{a}}{a} = -\frac{\kappa}{2}(\rho + 3(p+\Pi))\,, \tag{29.120}$$

and

$$\dot{\rho} + 3H(\rho + p + \Pi) = 0\,. \tag{29.121}$$

In the Eckart theory, the bulk viscosity pressure has the form

$$\Pi = -3\xi H\,. \tag{29.122}$$

In the Israel–Stewart theory the bulk viscosity pressure is given by the following differential equation:

$$\tau\dot{\Pi} + \Pi = -3\xi H - \frac{1}{2}\tau\Pi\left(3H + \frac{\dot{\tau}}{\tau} - \frac{\dot{\xi}}{\xi} - \frac{\dot{T}}{T}\right)\,, \tag{29.123}$$

where T is temperature and τ is relaxation time. Using (29.120) and the equation of state $p = w\rho = (\gamma - 1)\rho$, we obtain

$$\dot{H} = -\frac{3}{2}\gamma H^2 - \frac{\kappa}{2}\Pi\,. \tag{29.124}$$

In what follows we will solve (29.124) for Eckart and Israel–Stewart fluids. Inserting (29.119) and (29.122) into (29.124), yields

$$\dot{H} = \frac{9}{2}\xi_0 H^3 - \frac{3}{2}\gamma H^2 + \frac{3}{2}\kappa b H\,. \tag{29.125}$$

Integrating (29.125), we obtain implicitly $H(t)$ in the Eckart theory

$$t(H) = t_0 - \frac{2}{9\xi_0}\left[\ln\left(\frac{H}{H_0}\right)^{\frac{1}{\hat{H}^2-\left(\frac{\gamma}{6\xi_0}\right)^2}} + \ln\left(\frac{H_0 - \hat{H} - \frac{\gamma}{6\xi_0}}{H - \hat{H} - \frac{\gamma}{6\xi_0}}\right)^{\frac{1}{2\hat{H}\left(\hat{H}+\frac{\gamma}{6\xi_0}\right)}} + \ln\left(\frac{H_0 + \hat{H} - \frac{\gamma}{6\xi_0}}{H + \hat{H} - \frac{\gamma}{6\xi_0}}\right)^{\frac{1}{2\hat{H}\left(\hat{H}-\frac{\gamma}{6\xi_0}\right)}}\right]\,, \tag{29.126}$$

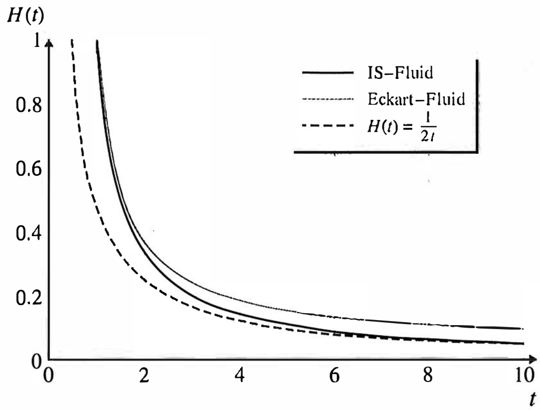


Fig. 29.2 The time evolution of the Hubble parameter from the Eckart theory, i. e., (29.126), Israel–Stewart theory in (29.131) and for radiation dominated universe without viscosity, i. e., $H(t) = 1/2t$. Here $b = 0.1$ and $\omega_0 = 1$

where

$$\hat{H}^2 = \left(\frac{\gamma}{6\xi_0}\right)^2 - \frac{b\kappa}{3\xi_0}\,.$$

If $\xi = \xi_0\rho$, i. e., $b = 0$, (29.126) reduces to

$$t(H) = t_0 - \frac{2}{3\xi_0} \times \left[\frac{1}{H} - \frac{1}{H_0} + \frac{3\xi_0}{\gamma}\ln\left(\frac{H/H - \frac{\gamma}{3\xi_0}}{H_0/H_0 - \frac{\gamma}{3\xi_0}}\right)\right]. \tag{29.127}$$

Assuming that the relaxation time is given by [29.7]

$$\tau = \xi\rho^{-1} \approx \xi_0\,, \tag{29.128}$$

and that the temperature has the form [29.44, 45]

$$T = \beta\rho^r\,, \tag{29.129}$$

where $\beta \approx 0.718$, $r \approx 0.213$, $\gamma \approx 1.183$, $\omega_0 \approx$ 0.5–1.5 GeV and

$$\xi_0 = \frac{1}{9\omega_0}\frac{9\gamma^2 - 24\gamma + 16}{\gamma - 1}\,, \tag{29.130}$$

we can use Israel–Stewart theory and insert (29.123) into (29.124) to get the equation that describes the cosmological evolution of the Hubble parameter

$$\ddot{H} + \frac{3}{2}[1 + (1-r)\gamma]H\dot{H} + \frac{1}{\alpha}\dot{H} - (1+r)\frac{\dot{H}^2}{H} + \frac{9}{4}(\gamma - 2)H^3 + \frac{3}{2}\frac{\gamma}{\alpha}H^2 = 0\,. \tag{29.131}$$

In the limit the viscosity vanishes, i. e., $\xi \to 0$, (29.131) reduces to

$$\dot{H} + \tfrac{3}{2}\gamma H^2 = 0\,. \tag{29.132}$$

Integrating this equation, we obtain

$$H(t) = \frac{2}{3\gamma}\frac{1}{t}\,. \tag{29.133}$$

In the radiation dominated era $\gamma = 4/3$, and (29.133) reduces to $H(t) = 1/2t$.

In Fig. 29.2 we have plotted the numerical solution of (29.131) for the time evolution of the Hubble parameter in the Israel–Stewart theory along with the corresponding equations from the Eckart theory, i. e., (29.126) and $H(t) = 1/2t$ for a radiation dominated era without viscosity.

29.9 Viscous Bianchi Type-I Universe Models

Bianchi type-I universe models are the simplest models of anisotropic universes that describe a homogeneous and spatially flat space-time and if filled with perfect fluid with the equation of state $p = w\rho$, $w < 1$, eventually evolve into a FRW universe. The isotropy of the present-day universe makes the Bianchi type-I models prime candidates for studying possible effects of an anisotropy in the early universe on modern-day observational data.

Some cosmologists have studied Bianchi type-I universe models with viscous fluid. The influence of viscosity on Bianchi type-I models has been investigated by *Belinskij* and *Khalatnikov* [29.46], and they found that asymptotically for large times, such Bianchi

type-I models will approach an isotropic steady-state universe model with a de Sitter space which expands exponentially. For asymptotically early times they found that there exists a Kasner era in which the effects of matter, radiation, and viscosity are negligible. Other authors [29.47, 48] have also concluded that anisotropic models have in general a vacuum stage near an unavoidable initial singularity in which the energy–momentum tensor has no influence on the cosmic evolution. But *Grøn* [29.5] has found that in a Bianchi type-I universe model filled with viscous Zel'dovich fluid, the bulk viscosity may remove the initial singularity. He also concluded that the viscosity and also LIVE [29.47], have an important role in isotropizing the universe.

We will, in this section, study the influence of viscosity on the evolution of homogeneous and anisotropic Bianchi type-I cosmological models, filled with nonlinear viscous fluid, both with and without a cosmological constant, Λ. In the present section, we generalize a recent analysis of viscous isotropic FRW-universe models [29.19] to anisotropic universe models.

The line element of a Bianchi type-I universe can be written in the form

$$\mathrm{d}s^2 = \mathrm{d}t^2 - R_i^2(\mathrm{d}x^i)^2\,, \tag{29.134}$$

where $R_1 = a(t)$, $R_2 = b(t)$, $R_3 = c(t)$ are the directional scale factors. The energy–momentum tensor of the viscous fluid has nonvanishing components

$$\begin{aligned} T_0^0 &= \rho\,, \\ T_i^i &= -p + 2\eta H_i + (3\xi - 2\eta)H - 9\alpha H \Delta H_i\,, \end{aligned} \tag{29.135}$$

where $H_i = \dot{R}_i/R_i$ are the directional Hubble parameters, $H = \frac{1}{3}\sum_{i=1}^{3} H_i$, $\Delta H_i = H_i - H$ and $p = w\rho$. For these universe models the Raychaudhuri equation takes the form [29.49]

$$\dot{H} = -3H^2 + \frac{\kappa}{2}(1-w)\rho + \frac{3}{2}\kappa\xi H + \Lambda\,. \tag{29.136}$$

The anisotropy parameter is defined as [29.50]

$$A = \frac{1}{3}\sum_{i=1}^{3}\left(\frac{\Delta H_i}{H}\right)^2 = \frac{1}{9}\sum_{i<j}\left(\frac{H_i - H_j}{H}\right)^2\,. \tag{29.137}$$

From Einstein's field equations then follow

$$\kappa\rho = \left(1 - \frac{A}{2}\right)3H^2 - \Lambda\,, \tag{29.138}$$

and

$$\begin{aligned} A &= C\frac{\tau^{2(3\alpha-1)}\mathrm{e}^{-2\Phi}}{9H^2}\,, \\ \tau &= abc\,, \\ \Phi &= 2\int \eta\,\mathrm{d}t\,, \end{aligned} \tag{29.139}$$

where C is an integration constant. Inserting (29.139) into (29.138) leads to

$$\kappa\rho = 3H^2 - \frac{C}{6}\tau^{2(3\alpha-1)}\mathrm{e}^{-2\Phi} - \Lambda\,. \tag{29.140}$$

Further inserting (29.140) into (29.136) gives

$$\begin{aligned} \dot{H} = &-\frac{3}{2}(1+w)H^2 + \frac{3}{2}\kappa\xi H \\ &- \frac{C}{12}(1-w)\tau^{2(3\alpha-1)}\mathrm{e}^{-2\Phi} + \frac{1}{2}(1+w)\Lambda\,. \end{aligned} \tag{29.141}$$

29.9.1 Bianchi Type-I Universe with Viscous Zel'dovich Fluid and LIVE

As a simple example showing some properties of an anisotropic, viscous universe model we shall first consider a Bianchi type-I universe model with a viscous fluid consisting of a mixture of a Zel'dovich fluid (also called stiff matter because the velocity of sound is equal to the velocity of light c in such a fluid) with $w = 1$ and LIVE with $w = -1$. Then (29.136) reduces to

$$\dot{H} = -3H^2 + \tfrac{3}{2}\kappa\xi H + \Lambda\,. \tag{29.142}$$

Assuming that the bulk viscosity is constant we get the following expression for the Hubble parameter:

$$H(t) = \frac{\kappa\xi}{4} + \hat{H}\coth\left(3\hat{H}t\right)\,, \quad \hat{H}^2 = \left(\frac{\kappa\xi}{4}\right)^2 + \frac{\Lambda}{3}\,. \tag{29.143}$$

The volume scale factor normalized to unity at the present time takes the form

$$\tau(t) = \mathrm{e}^{\frac{3\kappa\xi}{4}(t-t_0)}\frac{\sinh\left(3\hat{H}t\right)}{\sinh\left(3\hat{H}t_0\right)}\,. \tag{29.144}$$

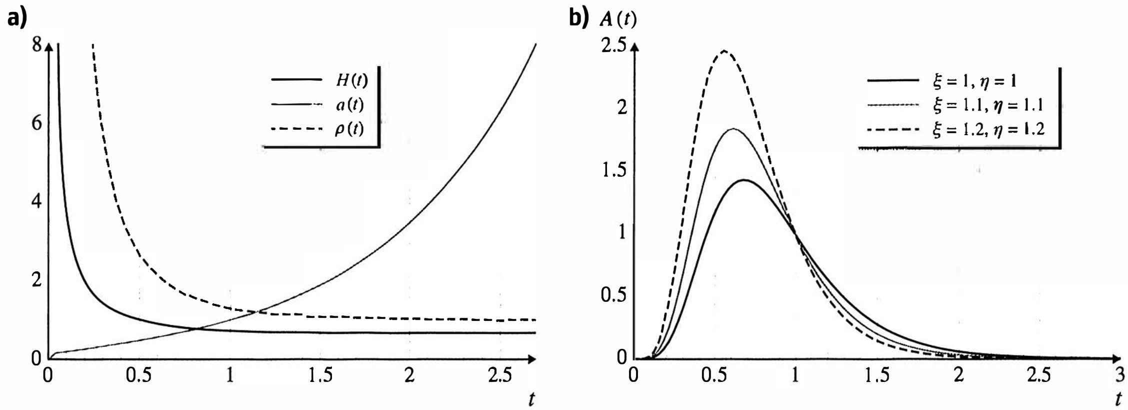


Fig.29.3a,b The Hubble parameter, the scale factor, the energy density, and the anisotropy parameter as functions of time. (**a**) $H(t)$, $\tau(t)$ and $\rho(t)$. Here $\alpha = 0.3$ and $\Lambda = 0.3$; (**b**) $A(t)$. Here $\alpha = 0.3$ and $\Lambda = 0.3$

The anisotropy parameter varies with time as

$$A(t) \propto \frac{e^{((3\kappa\xi/2)(3\alpha-1)-4\eta)t}}{\left[\frac{\kappa\xi}{4} + \hat{H}\coth\left(3\hat{H}t\right)\right]^2} . \tag{29.145}$$

The time evolution of the Hubble parameter, the scale factor, the energy density, and the anisotropy parameter are shown in Fig. 29.3. As we see from Fig. 29.3a the Hubble parameter and hence, from (29.140) also the density of the fluid, are both infinitely large at the beginning of the cosmic evolution. As t increases the Hubble parameter and the energy density decrease and approach finite values. The universe starts from a Big Bang with vanishing value of the volume scale factor. From Fig. 29.3b we see that the bigger the values of the shear and the bulk viscosities are the faster the anisotropy parameter goes to zero. This means that the shear and bulk viscosities contribute to isotropization of the universe. The presence of the shear and the bulk viscosities will also contribute to energy production of the universe.

29.9.2 Bianchi Type-I Universe with Variable Shear and Bulk Viscosity

In order to consider a simple example of an anisotropic universe model with both shear and bulk viscosity we shall assume that the coefficient of shear viscosity is proportional to the Hubble parameter H averaged over the different directions, with a carefully chosen proportionality constant,

$$\eta = -\tfrac{3}{2}(1-3\alpha)H . \tag{29.146}$$

so that $\tau^{2(3\alpha-1)}e^{-2\Phi} = 1$. Furthermore, we assume that the coefficient of bulk viscosity is given by

$$\xi = \xi_0 + \xi_1\frac{\dot{\tau}}{\tau} + \xi_2\frac{\ddot{\tau}}{\dot{\tau}} . \tag{29.147}$$

Then (29.141) reduces to

$$a\dot{H} = bH^2 + cH + d . \tag{29.148}$$

where

$$\begin{aligned} a &= 1 - \frac{3}{2}\kappa\xi_2 , \\ b &= \frac{3}{2}\left[3\kappa(\xi_1+\xi_2) - (1+w)\right] , \\ c &= \frac{3}{2}\kappa\xi_0 , \\ d &= \frac{1}{2}(1+w)\Lambda - \frac{C}{12}(1-w) . \end{aligned} \tag{29.149}$$

Integration with $\tau(0) = 0$, $\tau(t_0) = 1$ and assuming $\xi_1 + \xi_2 < 1/3$, gives

$$H(t) = -\frac{c}{b} + \hat{H}\coth\left(-\frac{b}{a}\hat{H}t\right) , \tag{29.150}$$

and

$$\tau(t) = e^{-\frac{3c}{2b}(t-t_0)}\left(\frac{\sinh\left(-\frac{b}{a}\hat{H}t\right)}{\sinh\left(-\frac{b}{a}\hat{H}t_0\right)}\right)^{-3b/a} , \tag{29.151}$$

where

$$\hat{H}^2 = \left(\frac{c}{2b}\right)^2 - \frac{d}{b}\,, \tag{29.152}$$

and

$$t_0 = -\frac{a}{b\hat{H}} \operatorname{artanh}\left(\frac{\hat{H}}{H_0 + \frac{c}{2b}}\right). \tag{29.153}$$

In this case the anisotropy parameter is given by

$$A = A_0 \left(\frac{H_0}{H^2}\right)^2, \tag{29.154}$$

and the density is

$$\rho = \rho_0 + 3\left(H^2 - H_0^2\right). \tag{29.155}$$

For pressureless matter with $w = 0$, we find that in this model the universe starts with a big bang at $t = 0$ with zero anisotropy. As t increases the volume of this universe increases, but the energy density decreases. The energy density decreases faster for smaller values of bulk viscosity, which means that the bulk viscosity plays an important role in the energy production of the universe. The anisotropy parameter increases with time, but it will approach a finite value. The bigger the value of the bulk viscosity is the smaller is the anisotropy parameter. This means that the viscosity contributes in keeping the anisotropy of the universe small.

29.9.3 Decaying Vacuum Energy

Bali et al. [29.51] have considered a related universe model with Zel'dovich fluid, $w = 1$, and a decaying vacuum energy with density proportional to the Hubble parameter, $\Lambda = aH$, where a is a positive constant. Then (29.146) reduces to

$$\dot{H} = -\frac{3}{2}(2 - \xi_1)H^2 + \left(a + \frac{3}{2}\xi_0\right)H\,. \tag{29.156}$$

The general solution is

$$H = \frac{bH_0}{H_0 d - (H_0 d - b)\mathrm{e}^{-b(t-t_0)}}\,, \tag{29.157}$$

where $b = a + (3/2)\xi_0$, $d = 3 - (3/2)\xi_1$ and $H_0 = H(t_0)$. The initial value of the Hubble parameter is

$$H(0) = \frac{bH_0}{H_0 d - (H_0 d - b)\mathrm{e}^{bt_0}}\,. \tag{29.158}$$

Considering a universe model with an initial Big Bang having $H(0) = \infty$ gives the age of the universe model in terms of the present value of the Hubble parameter,

$$t_0 = -\frac{1}{b}\ln\left(1 - \frac{b}{H_0 d}\right). \tag{29.159}$$

For this universe model the expression for the Hubble parameter reduces to

$$H = \frac{b}{d(1 - \mathrm{e}^{-bt})}\,. \tag{29.160}$$

Introducing an average scale factor $R = (R_1 R_2 R_3)^{1/3}$ so that $H = \frac{\dot{R}}{R}$ and integrating with the normalization $R(t_0) = 1$ we obtain

$$R = \left(\frac{\mathrm{e}^{bt} - 1}{\mathrm{e}^{bt_0} - 1}\right)^{1/d}. \tag{29.161}$$

The decay of the density of the vacuum energy is given by

$$\Lambda = \frac{ab}{d(1 - \mathrm{e}^{-bt})}\,. \tag{29.162}$$

The deceleration parameter as given in (29.5), is

$$q = d\mathrm{e}^{-bt} - 1\,. \tag{29.163}$$

At early times $q \approx 2 - (3/3)\xi_1 > 0$, and the expansion decelerates. The deceleration is reduced by the component of the bulk viscosity proportional to the Hubble parameter. At a point of time t_1 given by $q(t_1) = 0$, i.e., at $t_1 = (1/b)\ln d$ there is a transition from decelerated to accelerated expansion.

In the limit of large times this universe model enters a de Sitter era with the constant Hubble parameter

$$H(t \to \infty) = \frac{a + \frac{3}{2}\xi_0}{3 - \frac{3}{2}\xi_1}\,, \tag{29.164}$$

constant density of the vacuum energy, and constant deceleration parameter $q(t \to \infty) = -1$.

29.9.4 Anisotropic Bianchi Type-I Viscous Universe Models with Variable *G* and *Λ*

We have already considered isotropic universe models with variable G and Λ. It is time to look at anisotropic

Bianchi type-I viscous universe models with variable G and Λ. Many authors (see [29.52–56] and references therein) have studied different universe models in the presence of perfect or imperfect fluid with variable G and Λ, within the framework of Einstein's theory of relativity. We will here write down the dynamical equations for the Bianchi type-I universe models in the presence of nonlinear viscous fluid with variable G and Λ, and give a special solution.

In a comoving reference frame with diagonal metric tensor the equation of energy conservation $T^{\nu}_{0;\nu} = 0$ may be written as

$$\dot{T}^0_0 + (\ln\sqrt{-g})^{\cdot} T^0_0 - \tfrac{1}{2}\dot{g}_{\alpha\alpha}T^{\alpha\alpha} = 0\,. \tag{29.165}$$

Inserting the components of the energy momentum tensor in (29.135), we obtain

$$\dot{\rho} + 3H(\rho + p) = 3(3\xi - 2\eta)H^2 + 2\eta\sum_{i=1}^{3} H_i^2 + 27\alpha H^3 - 9\alpha H\sum_{i=1}^{3} H_i^2\,. \tag{29.166}$$

Using the definition of the anisotropy parameter, we can rewrite this equation as

$$\dot{\rho} + 3H(\rho + p) = 3(3\xi + 2\eta A)H^2 - 27\alpha A H^3\,. \tag{29.167}$$

From the Bianchi identities we obtain

$$\dot{\rho} + 3H(\rho + p) - 3(3\xi + 2\eta A)H^2 + 27\alpha A H^3 = -\frac{\dot{G}}{G}\rho - \frac{\dot{\Lambda}}{8\pi G}\,. \tag{29.168}$$

From (29.167) and (29.168), we have

$$\frac{\dot{G}}{G}\rho + \frac{\dot{\Lambda}}{8\pi G} = 0\,. \tag{29.169}$$

From (29.139)–(29.141) and (29.167) and (29.169) it follows that we have four independent equations having nine unknowns w, ξ, η, α, τ, H, Λ, G, and ρ. In what follows we give a solution to these equation based on the work of [29.52].

We start by assuming that $\alpha = \eta = 0$, and

$$b(t) = c(t)\,. \tag{29.170}$$

With this assumption we get

$$3H(t) = \frac{\dot{\tau}}{\tau} = \frac{\dot{a}}{a} + 2\frac{\dot{b}}{b}\,. \tag{29.171}$$

Furthermore, since the majority of the universe models has a scale factor that is either given by a power law form or exponential form, we assume

$$a(t) = a_0 t^n\,,\quad b(t) = b_0 t^l\,, \tag{29.172}$$

where a_0, b_0, n, and l are constants. We also assume

$$\Lambda = \Lambda_0\left(\frac{\dot{a}}{a} + 2\frac{\dot{b}}{b}\right)^2, \tag{29.173}$$

where Λ_0 is a constant. Using the above assumptions we obtain the following solutions:

$$\Lambda(t) = \frac{m}{t^2}, \tag{29.174}$$

$$H(t) = \frac{n + 2l}{3t}, \tag{29.175}$$

$$G(t) = G_0\left(\frac{t}{t_0}\right)^{\frac{2m}{k-m}}, \tag{29.176}$$

$$\rho(t) = \rho_0 t^{\frac{-2k}{k-m}}, \tag{29.177}$$

$$A(t) = 2\left[1 - 3\Lambda_0 - 3(n + 2l)^2\rho_0 t^{\frac{-2m}{k-m}}\right], \tag{29.178}$$

$$\xi(t) = \xi_0 t^{-\frac{k+m}{k-m}}, \tag{29.179}$$

where G_0 is the value of the gravitational constant at the precent time t_0, and $k = l^2 + 2nl$, $m = \Lambda_0(n + 2l)^2$, and

$$\rho_0 = \frac{k-m}{8\pi G_0} t_0^{\frac{2m}{k-m}}, \tag{29.180}$$

$$\xi_0 = \frac{\rho_0\Lambda_0}{m}\left[\gamma(n + 2l) - \frac{2k}{k-m}\right]. \tag{29.181}$$

Singh and *Kale* [29.52] have considered the case where $n = l$, and by using the observational values of $\frac{\dot{G}}{G} \approx 10^{-11}$ and $t_0 \approx 10^{10}$ they have found that the deceleration parameter is $q = \frac{1-n}{n} = -0.1150$, which is within the observational limits. They have also found that for $n > 1$ this model describes a universe with accelerated expansion.

29.10 Viscous Cosmology with Casual Thermodynamics

So far we have reviewed cosmological models of the universe with nonequilibrium thermodynamical processes mostly described by the theories of *Eckart* [29.3] and *Landau* and *Lifshitz* [29.57]. But the Eckart formalism is not completely consistent because it is restricted to a first-order deviation from equilibrium and therefore suffers from serious drawbacks concerning stability and causality. We will therefore look at homogenous and isotropic universe models with the causal second-order theories of nonequilibrium thermodynamical processes due to the work of *Müller* [29.58], *Israel* [29.59], *Israel* and *Stewart* [29.60], *Pavón* et al. [29.61] and *Hiscock* and *Lindblom* [29.62].

Belinskii et al. [29.63] were the first to study the cosmological implications of Müller–Israel–Stewarts theory, followed by *Pavón* et al. [29.61], *Grøn* [29.5], *Maartens* [29.6, 7], and *Zimdahl* [29.8, 64]. For the impact of the bulk viscosity on the background expansion of the universe the reader is referred to [29.65–74]. For the perturbative analysis of the viscous cosmological models we cite the following papers [29.28], [29.29], and [29.75–79]. Reheating and causal thermodynamics have been discussed by *Zimdahl* et al. [29.80].

In what follows, we will look at two causal bulk viscous cosmological models of the universe, one with conservation of the fluid particle number and the other without this conservation law, i. e., with particle production.

29.10.1 Causal Bulk Viscosity with Particle Conservation

In this section, we will study the viscous universe models by using the Israel–Stewart theory of causal thermodynamics. We start with the entropy flow vector S^{μ}, which with second-order deviation from equilibrium takes the form

$$S^{\mu} = sN^{\mu} - \frac{\tau\Pi}{2\xi T}u^{\mu}\,, \tag{29.182}$$

where u^{μ} is the four-velocity, T is the temperature, ξ is the coefficient of bulk viscosity, τ is the relaxation time, s is the entropy per particle, Π is the bulk viscous pressure, and $N^{\mu} = nu^{\mu}$ is particle flow vector, where n is the particle number density. Particle conservation $N^{\mu}_{;\mu} = 0$ and energy–momentum conservation $T^{\mu}_{;\mu} = 0$ imply

$$\dot{n} + \Theta n = 0\,, \tag{29.183}$$

and

$$\dot{\rho} = -\Theta(\rho + p + \Pi)\,, \tag{29.184}$$

respectively. From (29.182) and (29.183), we obtain the following expression for the entropy production density:

$$S^{\mu}_{;\mu} = -\frac{\Pi}{T}\left[\Theta + \frac{\tau}{\xi}\dot{\Pi} + \frac{1}{2}\Pi T\left(\frac{\tau}{\xi T}u^{\mu}\right)_{;\mu}\right]. \tag{29.185}$$

In order to satisfy the second law of thermodynamics, i. e., $S^{\mu}_{;\mu} \geq 0$, we have to impose the linear relation

$$\Pi = -\xi\left[\Theta + \frac{\tau}{\xi}\dot{\Pi} + \frac{1}{2}\Pi T\left(\frac{\tau}{\xi T}u^{\mu}\right)_{;\mu}\right], \tag{29.186}$$

which leads to

$$\tau\dot{\Pi} + \Pi = -\xi\Theta - \frac{1}{2}\xi\Pi T\left(\frac{\tau}{\xi T}u^{\mu}\right)_{;\mu}. \tag{29.187}$$

By using that $u^{\mu}_{;\mu} = \Theta = 3H$, $\dot{n} = n_{,\mu}u^{\mu}$, $\dot{T} = T_{,\mu}u^{\mu}$ and $\dot{\xi} = \xi_{,\mu}u^{\mu}$, (29.187) takes the form

$$\tau\dot{\Pi} + \Pi = -3\xi H - \frac{1}{2}\tau\Pi\left(3H + \frac{\dot{\tau}}{\tau} - \frac{\dot{\xi}}{\xi} - \frac{\dot{T}}{T}\right). \tag{29.188}$$

This dynamical equation determines the evolution of the viscous pressure. When the relaxation time vanishes, i. e., $\tau \to 0$, (29.188) reduces to

$$\Pi = -3\xi H\,, \tag{29.189}$$

which is the corresponding relation of the Eckart theory. When we assume that the second term on the right hand of (29.187) is negligible compared with the other terms in the equation, we obtain (what *R. Maartens* [29.7] calls) the truncated Israel–Stewart equation, which is of covariant relativistic Maxwell–Cattaneo form

$$\tau\dot{\Pi} + \Pi = -3\xi H\,. \tag{29.190}$$

The explicit criteria for using this equation over the full Israel–Stewart equation is given by *Zimdahl* [29.8]. We assume the following general form for the equations of state:

$$p = p(n, T) \quad \text{and} \quad \rho = \rho(n, T). \tag{29.191}$$

The temperature of the viscous fluid is determined by [29.7, 8]

$$\frac{\dot{T}}{T} = -3H\left[\left(\frac{\partial p}{\partial \rho}\right)_n + \frac{\Pi}{T}\left(\frac{\partial T}{\partial \rho}\right)_n\right]. \tag{29.192}$$

For $\Pi = 0$ and for the equation of state $p = (\gamma - 1)\rho$, (29.192) reduces to

$$\frac{\mathrm{d}T}{T} = -3(\gamma - 1)\frac{\mathrm{d}a}{a}, \tag{29.193}$$

which after integration gives $T \propto a^{-1}$, for $\gamma = \frac{4}{3}$, i. e. for a radiation dominated universe and $T \propto a^{-2}$, for $\gamma = 1$, i. e. for a matter dominated universe. The bulk viscous pressure is expected to be negative, therefore, from (29.192) we see that in the presence of bulk viscosity the temperature decreases less. Using (29.191) and (29.192), we can rewrite (29.188) as

$$\begin{aligned} \tau\dot{\Pi} + \Pi = {} & -3\tau\rho H\left[\gamma c_\mathrm{b}^2 + \frac{\Pi}{2\rho}\left(2 + \frac{\partial p}{\partial \rho} + c_\mathrm{s}^2\right)\right. \\ & \left. + \frac{\Pi^2}{2\gamma\rho^2}\left(1 + \frac{\partial p}{\partial \rho} + \frac{\rho + p}{T}\frac{\partial T}{\partial \rho}\right)\right] \\ & + \frac{\tau\Pi}{2}\frac{(\dot{c_\mathrm{b}^2})}{c_\mathrm{b}^2}, \end{aligned} \tag{29.194}$$

where

$$c_\mathrm{s}^2 = \left(\frac{\partial p}{\partial \rho}\right)_{ad} = \frac{n}{\rho + p}\frac{\partial p}{\partial n} + \frac{T}{\rho + p}\frac{(\partial p/\partial T)^2}{\partial \rho/\partial T} \tag{29.195}$$

is the square of the adiabatic sound velocity, c_S, and

$$c_\mathrm{b}^2 = \frac{\xi}{(\rho + p)\tau} \tag{29.196}$$

is the speed of bulk viscous perturbations, which is the nonadiabatic contribution of the speed of sound v given by

$$v^2 = c_\mathrm{b}^2 + c_\mathrm{s}^2 \le 1 \tag{29.197}$$

in a dissipative fluid without heat flux or shear viscosity. When we have the equation of state $p = (\gamma - 1)\rho$, (29.195) gives $c_\mathrm{s}^2 = \gamma - 1$, so that

$$c_\mathrm{b}^2 \le 2 - \gamma\,. \tag{29.198}$$

For a flat, homogeneous and isotropic universe, Einstein's field equations give

$$H^2 = \frac{\kappa}{3}\rho\,, \tag{29.199}$$

$$\frac{\ddot{a}}{a} = -\frac{\kappa}{6}\left[\rho + 3(p + \Pi)\right]. \tag{29.200}$$

From (29.200), we obtain

$$\kappa\Pi = -2\dot{H} - 3\gamma H^2\,. \tag{29.201}$$

Differentiating (29.201), we get

$$\kappa\dot{\Pi} = -2\ddot{H} - 6H\dot{H}\left(1 + \frac{\partial p}{\partial \rho}\right) + 9\gamma H^3\left(c_\mathrm{s}^2 - \frac{\partial p}{\partial \rho}\right). \tag{29.202}$$

Inserting (29.201) and (29.202) into (29.194), we obtain a dynamical equation for the causal evolution of the Hubble parameter

$$\begin{aligned} \tau H & \left[\frac{\ddot{H}}{H} - A\frac{\dot{H}^2}{H^2} - 3\dot{H}\left(B + \frac{1}{2}C\right)\right. \\ & - \frac{9}{2}\gamma H^2\left(c_\mathrm{b}^2 + \frac{1}{2}(B - C)\right) \\ & \left. - \frac{1}{2}\frac{(\dot{c_\mathrm{b}^2})}{Hc_\mathrm{b}^2}\left(\dot{H} + \frac{3}{2}\gamma H^2\right)\right] \\ + \dot{H} & + \frac{3}{2}\gamma H^2 = 0, \end{aligned} \tag{29.203}$$

where

$$\begin{aligned} r &= \frac{\rho + p}{T}\frac{\partial T}{\partial \rho}\,, \\ A &= \gamma^{-1}\left(r + 1 + \frac{\partial p}{\partial \rho}\right), \\ B &= r - 1 - \frac{\partial p}{\partial \rho}\,, \\ C &= \frac{\partial p}{\partial \rho} - c_\mathrm{s}^2\,. \end{aligned} \tag{29.204}$$

For nonrelativistic matter $\gamma = 1$, $\partial p/\partial \rho = 2/3$, and $c_\mathrm{s}^2 \ll 1$, which implies that $r \gg 1$, $A \gg 1$, $B \gg 1$ and $C = 2/3$. For radiation $\gamma = 4/3$, $\partial p/\partial \rho = c_\mathrm{s}^2 = 1/3$, therefore, $r = 1, A = 2$ and $B = C = 0$ in that case.

Stationary Solutions, i. e., $\dot{H}=0$

Assuming that c_b is a constant and $H = H_0 = constant$, (29.203) gives

$$\tau H\left\{-\frac{9}{2}\gamma H^2\left[c_b^2+\frac{1}{2}(B-C)\right]\right\}+\frac{3}{2}\gamma H^2=0\,, \tag{29.205}$$

which we rewrite as

$$\tau H=\frac{1}{3\left[c_b^2+\frac{1}{2}(B-C)\right]}\,. \tag{29.206}$$

For nonrelativistic matter we have $c_b^2 \leq 1$, $B \gg 1$ and $C \approx 2/3$, (29.206) yields

$$\tau H \ll 1 \quad \Rightarrow \quad \tau \ll \frac{1}{H}\,. \tag{29.207}$$

This means that the relaxation time is much shorter than the cosmological time scale H^{-1}. For radiation $c_b^2 \leq 2/3$, $r = 1$ and $B = C = 0$, (29.206) gives

$$\tau H=\frac{1}{3c_b^2}\geq\frac{1}{2} \quad \Rightarrow \quad \tau\geq\frac{1}{2H}\,. \tag{29.208}$$

In this case the relaxation time may well be of order of the cosmological time H^{-1}, and the nonequilibrium is said to be *frozen in*. It is a necessary condition for a successful inflation that the relaxation time is of the same order as the Hubble time. Since the Hubble parameter is a constant, the scale factor is $a \propto \exp(H_0 t)$. From (29.199), (29.200), and (29.201) we obtain

$$\Pi=-\frac{3\gamma}{\kappa}H_0^2=-\gamma\rho\,. \tag{29.209}$$

Using (29.192) for the evolution of the temperature and (29.183) for particle conservation, we find

$$T\propto a^{3\left(r-\frac{\partial p}{\partial\rho}\right)}\,, \tag{29.210}$$

$$n\propto a^{-3}\,, \tag{29.211}$$

respectively. These equations imply that the temperature is exponentially increasing, but the number density is exponentially decreasing, which results in a constant energy density. With these solutions *Zimdahl* [29.8] concludes that this evolution is unrealistic.

Power Law Solutions, i. e., $a \propto t^q$

With a power law solution, i. e., $a \propto t^q$, the Hubble parameter takes the form $H = \frac{\dot{a}}{a} = qt^{-1}$. Inserting

$$\dot{H}=-qt^{-2} \quad \text{and} \quad \ddot{H}=2qt^{-3} \tag{29.212}$$

into (29.203), gives

$$\tau H=\frac{1-\frac{2}{3\gamma q}}{3c_b^2}Q\,, \tag{29.213}$$

where

$$Q=\frac{c_b^2}{c_b^2+\frac{1}{2}(B-C)+\frac{2}{9\gamma q^2}\left[A-2-3q(B+C)\right]}\,. \tag{29.214}$$

For nonrelativistic matter $Q \ll 1$, and therefore, the relaxation time is much shorter than the Hubble time, i. e., $\tau \ll H^{-1}$. For ultrarelativistic particles $Q \to 1$, which implies that

$$\tau\geq\left(\frac{1}{2}-\frac{1}{3q}\right)\frac{1}{H}\,. \tag{29.215}$$

This means that in this case the relaxation time may well be of the order of the Hubble time. Since the relaxation time τ and the bulk viscous coefficient ξ are positively defined and $Q > 0$ the bulk pressure must be negative. From (29.199) and (29.200), we obtain

$$-\frac{\Pi}{\rho+p}=1-\frac{2}{3\gamma q}\,. \tag{29.216}$$

Combining (29.213) and (29.216), we get

$$\tau H=-\frac{1}{3c_b^2}\frac{\Pi}{\rho+p}Q\,. \tag{29.217}$$

From this equation we see that when $\Pi \neq 0$ in the case of the Eckart theory we have $\tau \to 0$ in the limit $c_b^2 \to \infty$. Solving (29.216) for q gives

$$q=\frac{2}{3\gamma\left(1+\frac{\Pi}{\rho+p}\right)}\,. \tag{29.218}$$

The evolution of the temperature is obtained from (29.192)

$$T\propto a^{3\left[r\left(1-\frac{2}{3\gamma q}\right)-\frac{\partial p}{\partial\rho}\right]}\,. \tag{29.219}$$

Equation (29.218) implies that $q = 2/(3\gamma)$ corresponds to $\Pi = 0$, which is the limit for the perfect fluid. The number density takes the form $n \propto a^{-3} \propto t^{-3q}$, which means that it decays for $q > 0$, n is constant for $q = 0$ and it increases for $q < 0$. The case $q = 1$ gives the following solutions:

$$\begin{aligned} &a \propto t \quad \Rightarrow \quad H = t^{-1} \quad \Rightarrow \quad \rho = \frac{\kappa}{3} a^{-2}, \\ &\Pi = -\left(1 - \frac{2}{3\gamma}\right)(\rho + p) \\ &T \propto a^{3\left[r\left(1 - \frac{2}{3\gamma}\right) - \frac{\partial p}{\partial \rho}\right]} . \end{aligned} \tag{29.220}$$

These solutions describe a universe where the temperature is increasing but the energy density decreases. In order to have an inflationary behavior there must be accelerated expansion, $q > 1$. With these solutions, *Zimdahl* [29.64], concludes that *viscosity-driven inflation is hardly convincing.*

29.10.2 Causal Bulk Viscosity Without Particle Conservation

Assuming that the fluid particle number is not conserved in the early universe, i. e., $N^{\mu}_{;\mu} \neq 0$, we rewrite (29.183) as [29.64, 81, 82]

$$N^{\mu}_{;\mu} = \dot{n} + 3Hn = n\Gamma , \tag{29.221}$$

where $\Gamma = \frac{\dot{N}}{N}$ is the rate of change of the number $N = na^3$ of particles in a comoving volume a^3. For $\Gamma > 0$ we have particle creation, for $\Gamma < 0$ particles are annihilated. Inserting the continuity equations (29.184) and (29.221) into the Gibbs equation

$$T\,\mathrm{d}s = \mathrm{d}\left(\frac{\rho}{n}\right) + p\,\mathrm{d}\left(\frac{1}{n}\right), \tag{29.222}$$

we obtain

$$nT\dot{s} = -3H\Pi - (\rho + p)\Gamma . \tag{29.223}$$

The corresponding expression for the entropy production in that case is given by [29.64]

$$S^{\mu}_{;\mu} = n\dot{s} + sN^{\mu}_{;\mu} - \frac{1}{2}\left(\frac{\tau}{\xi T}u^{\mu}\right)_{;\mu} \Pi^2 - \frac{\tau}{\xi T}\Pi\dot{\Pi} . \tag{29.224}$$

We assume that we have *isentropic* particle production [29.64], which is characterized by $\dot{s} = 0$. With this condition, (29.223) gives

$$\Pi = -(\rho + p)\frac{\Gamma}{3H} , \tag{29.225}$$

which means that the bulk pressure is given by the particle production rate, and (29.224) may be written as [29.64]

$$\begin{aligned} TS^{\mu}_{;\mu} = (\rho + p)\frac{\Gamma}{3H} &\times \left\{\frac{3nsTH}{\rho + p} - (\rho + p)\frac{\Gamma}{3H}\frac{\tau}{2\xi}\right. \\ &\times \left[\Gamma + 2\frac{\dot{\Gamma}}{3H}\frac{3H}{\Gamma} - (3H - \Gamma)\left(c_s^2 - \frac{\partial p}{\partial \rho}\right)\right. \\ &\left.\left. - \frac{\dot{(c_b^2)}}{c_b^2}\right]\right\} , \end{aligned} \tag{29.226}$$

where we have used that

$$\frac{\dot{n}}{n} = -(3H - \Gamma) , \quad \frac{\dot{T}}{T} = -(3H - \Gamma)\frac{\partial p}{\partial \rho} \tag{29.227}$$

and

$$\dot{\rho} = -(3H - \Gamma)(\rho + p) , \quad \dot{p} = -c_s^2(3H - \Gamma)(\rho + p) . \tag{29.228}$$

In order to satisfy the second law of thermodynamics, i. e., $S^{\mu}_{;\mu} \geq 0$, we impose the generalized linear relation

$$\begin{aligned} &(\rho + p)\frac{\Gamma}{3H} \\ &= \xi\left\{\frac{3nsTH}{\rho + p}\right. \\ &\quad - (\rho + p)\frac{\Gamma}{3H}\frac{\tau}{2\xi}\left[\Gamma + 2\frac{\dot{\Gamma}}{3H}\frac{3H}{\Gamma}\right. \\ &\quad - (3H - \Gamma)\left(c_s^2 - \frac{\partial p}{\partial \rho}\right) \\ &\quad \left.\left. - \frac{\dot{(c_b^2)}}{c_b^2}\right]\right\} , \end{aligned} \tag{29.229}$$

which leads to

$$\tau\left(\frac{\Gamma}{3H}\right)^{\cdot}+\frac{\Gamma}{3H} = 3H\tau\left[\frac{nsT}{\rho+p}c_b^2-\frac{1}{2}\left(1+c_s^2-\frac{\partial p}{\partial\rho}\right)\left(\frac{\Gamma}{3H}\right)^2 +\frac{1}{2}\left(c_s^2-\frac{\partial p}{\partial\rho}\right)\frac{\Gamma}{3H}+\frac{1}{2}\frac{\dot{(c_b^2)}}{3Hc_b^2}\frac{\Gamma}{3H}\right]. \tag{29.230}$$

Using (29.225), we rewrite (29.230) in terms of Π

$$\tau\dot{\Pi}+\Pi = -3H\tau\rho\left[\frac{nsT}{\rho+p}\gamma c_b^2+\frac{\Pi}{2\rho}\left(2+c_s^2+\frac{\partial p}{\partial\rho}\right) +\frac{\Pi^2}{2\rho^2}\left(1+c_s^2+\frac{\partial p}{\partial\rho}\right)\right] + \frac{\tau\Pi}{2}\frac{\dot{(c_b^2)}}{c_b^2}. \tag{29.231}$$

Combining (29.199), (29.201), and (29.225), we obtain

$$\frac{\Gamma}{3H}=1+\frac{2}{3\gamma}\frac{\dot{H}}{H^2}. \tag{29.232}$$

Substituting (29.232) into (29.230), we obtain an equation for the evolution of the Hubble parameter

$$\tau H\left[\frac{\ddot{H}}{H}-A\frac{\dot{H}^2}{\gamma H^2}+3B\dot{H}-\frac{9}{2}\gamma H^2\left(rc_b^2-\frac{1}{2}\right) -\frac{1}{2}\frac{\dot{(c_b^2)}}{Hc_b^2}\left(\dot{H}+\frac{3}{2}\gamma H^2\right)\right] +\dot{H}+\frac{3}{2}\gamma H^2=0, \tag{29.233}$$

where

$$r=\frac{nTs}{\rho+p},\quad A=\left(1+c_s^2+\frac{\partial p}{\partial\rho}\right),\quad B=1-\frac{1}{2}\left(\frac{\partial p}{\partial\rho}-c_s^2\right). \tag{29.234}$$

For relativistic matter with $\gamma=4/3$, $\partial p/\partial\rho=c_s^2=1/3$, $ns=(\rho+p)/T$ and with the constant c_b^2, (29.233) reduces to

$$\tau H\left[\frac{\ddot{H}}{H}-\frac{5}{4}\frac{\dot{H}^2}{H^2}+3\dot{H}-6H^2\left(c_b^2-\frac{1}{2}\right)\right] +\dot{H}+2H^2=0. \tag{29.235}$$

In what follows we will study different solutions of (29.235).

Stationary Solutions $\dot{H}=0$

Assuming that the Hubble parameter is constant, i. e., $H=H_0$, (29.235) gives

$$\tau H=\frac{1}{3\left(c_b^2-\frac{1}{2}\right)}. \tag{29.236}$$

From this relation we see that $c_b^2>1/2$ in order to give a physically meaningful solution. Since $c_s^2=1/3$, from (29.197) we get the general causality restriction $c_b^2\le 1-c_s^2\le 2/3$. This means that c_b^2 takes values in the interval $1/2<c_b^2\le 2/3$. With this causality restriction we find

$$2\le\tau H<\infty, \tag{29.237}$$

which means that particle production makes the relaxation time larger (The *relaxation* time here may refer to a different process which is explained in [29.64]). From (29.227), (29.228) and (29.232) we obtain $\Gamma=3H$, and the particle density number n, the temperature T, the energy density ρ, and the pressure p are constants in time.

Solutions for the Case $\Gamma\propto H$

By using the relation

$$\dot{H}=\frac{\mathrm{d}H}{\mathrm{d}a}\dot{a}=H'Ha, \tag{29.238}$$

where $H'=\frac{\mathrm{d}H}{\mathrm{d}a}$, we can rewrite (29.232) as

$$\frac{\mathrm{d}H}{H\left(\frac{\Gamma}{3H}-1\right)}=\frac{3\gamma}{2}\frac{\mathrm{d}a}{a}. \tag{29.239}$$

Assuming that the particle production rate, Γ, is proportional to the Hubble parameter, H, i. e., $\Gamma=\alpha 3H$, with α being a constant, (29.239) reduces to .

$$\frac{\mathrm{d}H}{H}=\frac{3\gamma(\alpha-1)}{2}\frac{\mathrm{d}a}{a}. \tag{29.240}$$

Integration of this equation gives the following expression for the Hubble parameter in terms of the scale factor:

$$H(a) = H_0 \left(\frac{a}{a_0}\right)^{\frac{3}{2}\gamma(\alpha-1)} . \quad (29.241)$$

Using the definition of the Hubble parameter $H = \frac{\dot{a}}{a}$, and integrating (29.241), we obtain

$$a(t) = \left(\frac{t}{t_0}\right)^{\frac{2}{3\gamma(1-\alpha)}} . \quad (29.242)$$

The Hubble parameter in terms of the cosmic time is then $H(t) = \frac{2}{3\gamma(1-\alpha)}\frac{1}{t}$. From this solution we see that a linear relation between Γ and H implies a power-law behavior of the scale factor. From (29.227) and (29.228), it follows that

$$n = n_0 \left(\frac{t}{t_0}\right)^{-\frac{2}{\gamma}} = n_0 a^{-3(1-\alpha)} , \quad (29.243a)$$

$$T = T_0 \left(\frac{t}{t_0}\right)^{\frac{2(1-\gamma)}{\gamma}} = T_0 a^{3(1-\alpha)(1-\gamma)} , \quad (29.243b)$$

$$\rho = \rho_0 \left(\frac{t}{t_0}\right)^{-2} = \rho_0 a^{-3\gamma(1-\alpha)} , \quad (29.243c)$$

with $\gamma = 4/3$, $a = \left(\frac{t}{t_0}\right)^q$, where $q \equiv \frac{1}{2(1-\alpha)}$. Inserting $H = q\frac{1}{t}$, $\dot{H} = -q\frac{1}{t^2}$ and $\ddot{H} = 2q\frac{1}{t^3}$ into (29.235), yields

$$\tau H = \frac{2}{3} \frac{q(q-\frac{1}{2})}{q\left[1 + 2q\left(c_b^2 - \frac{1}{2}\right)\right] - \frac{1}{4}} . \quad (29.244)$$

With $q = 1/2$ we get $\Gamma = 0$ and $\tau = 0$, which means that the fluid is perfect. For $q > 1/2$ we have $\Gamma > 0$, and if we compare the parameters with those in (29.213) for $\gamma = 4/3$ and $Q = 1$, we find that the relaxation time is generally larger in this case.

Solutions for the Case $\Gamma \propto H^2$

To find yet another solution, we assume the following proportionality relation between the particle production rate Γ and H^2:

$$\frac{\Gamma}{3H} = \beta \frac{H}{H_e} , \quad (29.245)$$

where β is a constant and $H_e = H(a_e)$ is the Hubble parameter at some fixed epoch with $a = a_e$. Inserting (29.245) into (29.239) gives

$$\frac{dH}{H\left(\frac{\beta}{H_e}H - 1\right)} = \frac{3\gamma}{2}\frac{da}{a} . \quad (29.246)$$

Integrating this equation, we obtain

$$H(a) = H_e \frac{a_e^{\frac{3}{2}\gamma}}{(1-\beta)a^{\frac{3}{2}\gamma} + \beta a_e^{\frac{3}{2}\gamma}} . \quad (29.247)$$

From this expression we see that for $a \to 0$ we have $H \to \frac{H_e}{\beta} = constant$, which means that we have accelerated expansion ($\ddot{a} > 0$). For $a \gg a_e$ we have $H \propto a^{-3\gamma/2}$, which is the Hubble parameter for the FLRW universe. Assuming that $a = a_e$ in the transition from accelerated to decelerated expansion at which $\ddot{a} = 0$ and $\dot{H}_e = -H_e^2$, and combining (29.245) and (29.232), we find

$$\beta = 1 - \frac{2}{3\gamma} . \quad (29.248)$$

For radiation $\beta = 1/2$, and (29.247) reduces to

$$H(a) = \frac{2a_e^2 H_e}{a^2 + a_e^2} . \quad (29.249)$$

With this expression for the Hubble parameter the scale factor takes the form

$$t = t_e + \frac{1}{4H_e}\left[\ln\left(\frac{a}{a_e}\right)^2 + \left(\frac{a}{a_e}\right)^2 - 1\right] . \quad (29.250)$$

Inserting (29.249) in the first Friedmann equation, leads to the expression for the energy density

$$\rho(a) = \frac{12H_e^2}{\kappa}\left[\frac{a_e^2}{a^2 + a_e^2}\right]^2 . \quad (29.251)$$

From this expression we see that for $a \to 0$ we have $\rho \to \frac{12H_e^2}{\kappa} = constant$, and for $a \gg a_e$ we have the familiar behavior $\rho \propto a^{-4}$. Rewriting the dynamical equations for the particle number density and the temperature from (29.227), we obtain

$$\frac{dn}{n} = \frac{1}{3a}\left(\frac{a_e^2}{a^2 + a_e^2} - 1\right) \quad \text{and}$$

$$\frac{dT}{T} = \frac{1}{a}\left(\frac{a_e^2}{a^2 + a_e^2} - 1\right) , \quad (29.252)$$

where we have used the relation

$$\frac{\mathrm{d}}{\mathrm{d}t} = aH\frac{\mathrm{d}}{\mathrm{d}a} . \tag{29.253}$$

Integration of equations (29.252) yields

$$n = n_\mathrm{e}\left(\frac{2a_\mathrm{e}^2}{a^2 + a_\mathrm{e}^2}\right)^{\frac{3}{2}} \quad \text{and} \quad T = T_\mathrm{e}\left(\frac{2a_\mathrm{e}^2}{a^2 + a_\mathrm{e}^2}\right)^{\frac{1}{2}} . \tag{29.254}$$

These expressions show that both the particle number density and the temperature remain finite at $a = 0$, and they approach the correct relations for the radiation dominated era for $a \gg a_\mathrm{e}$. Inserting (29.249) into (29.235), gives

$$\tau H = \frac{2a_\mathrm{e}^2\left(a^2 + a_\mathrm{e}^2\right)}{3a^2 + 10a_\mathrm{e}^2 a^2 + 6\left(c_\mathrm{b}^2 - \frac{1}{2}\right)\left(a^2 + a_\mathrm{e}^2\right)^2} , \tag{29.255a}$$

$$= \frac{1}{6}\frac{1}{c_\mathrm{b}^2 + \frac{1}{3}\frac{H}{H_\mathrm{e}}\left(1 - \frac{7}{8}\frac{H}{H_\mathrm{e}}\right)}\frac{H}{H_\mathrm{e}} . \tag{29.255b}$$

In the limit $a \to 0$ we have $H \to 2H_\mathrm{e}$ and $\tau H = 1/[3(c_\mathrm{b}^2 - 1/2)]$. Since $1/2 < c_\mathrm{b}^2 \le 2/3$, we find $\tau = 2H^{-1}$, for the maximum value of c_b^2, i. e., for $c_\mathrm{b}^2 = 2/3$. From (29.255a) we also see that the parameter τH decreases with time, and for $a \gg a_\mathrm{e}$, τH tends to zero. Hence, in this limit we will have perfect fluid behavior, and since $H_\mathrm{e} \gg H$, the particle production is negligible, i. e., $\Gamma/H \ll 1$.

29.11 Summary

In this chapter, we have reviewed many viscous universe models. The noncausal first-order Eckart theory and the causal second-order Müller–Israel–Stewart (MIS) theory of dissipative processes in relativistic fluids have been applied to a flat, homogeneous and isotropic universe and the homogeneous but anisotropic Bianchi type-I universe. The generalization of the ΛCDM universe model to include viscosity, and viscous universe models with variable G and Λ have also been reviewed.

The possibility of an accelerated expansion of a matter dominated universe with constant bulk viscosity have been studied, and the estimated value of the density parameter of the viscous fluid that gives the best fit with observed data is $\Omega_{\xi 0} = 0.64$, which is 11 orders of magnitude greater than the value coming from kinetic gas theory. It is believed that a mechanism for producing greater viscosity may be generation of bulk viscosity due to decay of dark matter particles into relativistic products.

By using a special relation between the bulk viscosity and the Hubble parameter and its derivatives, and by assuming that the shear viscosity is proportional to the Hubble parameter, analytical solutions to the Raychaudhuri and the continuity equation have been found for the Bianchi type-I universe models with nonlinear viscous fluid. From these solutions it is clear that for these models the presence of bulk viscosity and the nonlinear viscous fluid will increase the energy density of matter. The evolution of the Hubble parameter, the volume scale factor, and the anisotropy parameter will also depend on the bulk viscosity. If the value of the bulk viscosity is increased the anisotropy parameter will decrease. In other words, the presence of viscosity is important for the isotropization of the universe.

The MIS theory with and without particle number conservation has been used to study the possibility of bulk viscosity-driven inflationary solutions. When the particle number is not conserved the bulk viscous pressure is assumed to be responsible for the particle production processes. In this case with a particle production rate which depends quadratically on the Hubble parameter a smooth transition from inflationary to noninflationary behavior was obtained.

References

29.1 C.W. Misner: Neutrino viscosity and the isotropy of primordial black body radiation, Phys. Rev. Lett. **19**, 533–535 (1967)

29.2 Y.B. Zel'dovoch, I.D. Novikov: *Relativistic Astrophysics*, Vol. 2 (Univ. Chicago Press, Chicago 1983)

29.3 C. Eckart: The Thermodynamics of irreversible processes. III. Relativistic theory of the simple fluid, Phys. Rev. **58**, 919–924 (1940)

29.4 W. Israel, J.M. Stewart: Thermodynamics of nonstationary and transients effects in a relativistic gas, Phys. Lett. **58**, 213–215 (1976)

29.5 Ø. Grøn: Viscous inflationary universe models, Astrophys. Space Sci. **173**, 191–225 (1990)

29.6 R. Maartens: Dissipative cosmology, Class. Quantum Gravity **12**, 1455–1466 (1995)

29.7 R. Maartens: Causal thermodynamics in relativity, Proc. Hanno Rund Workshop Relativ. Thermodyn. S. Afr. June 1996 (1996), arXiv: Astro-ph/9609119

29.8 W. Zimdahl: Bulk viscous cosmology, Phys. Rev. D **53**, 5483–5493 (1996)

29.9 M.K. Mak: Exact causal viscous cosmologies, Gen. Relativ. Gravit. **30**, 1171–1186 (1998)

29.10 B.C. Paul, S. Mukherjee, A. Beesham: Higher derivative theory with viscosity, Int. J. Mod. Phys. D **7**, 499–507 (1998)

29.11 I. Arbab, A. Beesham: Causal dissipative cosmology with variable G and Λ, Gen. Relativ. Gravit. **32**, 615–620 (2000)

29.12 S. Lepe, F. Peña, J. Saavedra: Randall–Sundrum model with $\Lambda < 0$ and bulk brane viscosity, Phys. Lett. B **662**, 217–219 (2008)

29.13 Y.B. Zel'dovich: The cosmological constant and the theory of elementary particles, Sov. Phys. Usp. **11**, 209–230 (1968)

29.14 Ø. Grøn: Repulsive gravitation and inflationary universe models, Am. J. Phys. **54**, 46–52 (1986)

29.15 Ø. Grøn: A new standard model of the universe, Eur. J. Phys. **23**, 135–144 (2002)

29.16 I.S. Kohli: A Bianchi type IV viscous fluid model of the early universe (2012), arXiv: astro-ph/1206.5438

29.17 J. Ren, X. Meng: Cosmological model with viscosity media (dark fluid) described by an effective equation of state, Phys. Lett. B **633**, 1–8 (2006)

29.18 M. Hu, X. Meng: Bulk viscous cosmology: Statefinder and entropy, Phys. Lett, B **635**, 186–194 (2006)

29.19 N. Mostafapoor, Ø. Grøn: Viscous ΛCDM universe models, Astrophys. Space Sci. **333**, 357–368 (2011)

29.20 I. Brevik, L.T. Heen: Remarks on the viscosity concept in the early universe, Astrophys. Space Sci. **219**, 99–115 (1994)

29.21 W. Hu, D.J. Eisenstein: The structure of structure formation theories, Phys. Rev. D **59**, 083509 (1999)

29.22 M. Kunz: The dark degeneracy: On the number and nature of dark components, Phys. Rev. D **80**, 123001 (2009)

29.23 X. Dou, X. Meng: Bulk viscous cosmology: unified dark matter, Commun. Theor. Phys. **52**, 377 (2010)

29.24 X. H. Meng, X. Dou: Singularity and entropy of the bulk viscosity dark energy model (2009), arXiv:astro-ph/0910.2397

29.25 X. Meng, Z. Ma: Rip/singularity free cosmology models with bulk viscosity, Eur. Phys. J. C **72**, 2053 (2012)

29.26 R. Amanullah, C. Lidman, D. Rubin, G. Aldering, P. Astier, K. Barbary, M.S. Burns, A. Conley, K.S. Dawson, S.E. Deustua, M. Doi, S. Fabbro, L. Faccioli, H.K. Fakhouri, G. Folatelli, A.S. Fruchter, H. Furusawa, G. Garavini, G. Goldhaber, A. Goobar, D.E. Groom, I. Hook, D.A. Howell, N. Kashikawa, A.G. Kim, R.A. Knop, M. Kowalski, E. Linder, J. Meyers, T. Morokuma, S. Nobili, J. Nordin, P.E. Nugent, L. Ostman, R. Pain, N. Panagia, S. Perlmutter, J. Raux, P. Ruiz-Lapuente, A.L. Spadafora, M. Strovink, N. Suzuki, L. Wang, W.M. Wood-Vasey, N. Yasuda: Spectra and Hubble telescope light curves of six type Ia supernovae at $0.511 < 1.12$ and the Union2 compilation, Astrophys. J. **716**, 712 (2010)

29.27 M. Moresco, L. Verde, L. Pozzetti, R. Jimenez, A. Cimatti: New constraints on cosmological parameters and neutrino properties using the expansion rate of the universe to $z \sim 1.75$, J. Cosmol. Astropart. Phys. **07**, 053 (2012)

29.28 W. Hipolito-Ricaldi, H. Velten, W. Zimdahl: Non-adiabatic dark fluid cosmology, J. Cosmol. Astropart. Phys. **0906**, 016 (2009)

29.29 R. Colistete, J. Fabris, J. Tossa, W. Zimdahl: Bulk viscous cosmology, Phys. Rev. D **76**, 103516 (2007)

29.30 M. Szydlowski, O. Hrycyna: Dissipative or conservative cosmology with dark energy?, Ann. Phys. **322**, 2745–2775 (2007)

29.31 A. Avelino, U. Ucamendi: Can a matter dominated model with constant bulk viscosity drive the accelerated expansion of the universe?, J. Cosmol. Astropart. Phys. **0904**, 006 (2009)

29.32 J.R. Wilson, G.J. Mathews, G.M. Fuller: Bulk viscosity, decaying dark matter and the cosmic acceleration, Phys. Rev. D **75**, 043521 (2007)

29.33 G.J. Mathews: Late decaying dark matter, bulk viscosity and the cosmic acceleration, Phys. Rev. D **78**, 043525 (2008)

29.34 J.A. Belinchón, T. Harko, M.K. Mak: Full causal bulk viscous cosmological models with variable G and Λ, Gravit. Cosmol. **8**, 319–326 (2002)

29.35 M.K. Mak, J.A. Belinchón, T. Harko: Causal bulk viscous dissipative isotropic cosmologies with variable gravitational and cosmological constants, Int. J. Mod. Phys. D **11**, 1265–1284 (2002)

29.36 P.A.M. Dirac: A new basis for cosmology, Proc. R. Soc. A **165**, 199–208 (1938)

29.37 A.-M.M. Abdel-Rahman: Singularity-free decaying-vacuum cosmologies, Phys. Rev. D **45**, 3497–3511 (1992)

29.38 J.D. Barrow, P. Parsons: Behavior of cosmological models with varying G, Phys. Rev. D **55**, 1906–1936 (1997)

29.39 A. Beesham: Cosmological models with a variable cosmological term and bulk viscous models, Phys. Rev. D **48**, 3539–3543 (1993)

29.40 A. Abdussatar, R.G. Vishwakarma: Some FRW models with variable G and Λ, Class. Quantum Gravity **14**, 945–953 (1997)

29.41 V. Mendez, D. Pavón: Expanding models with a varying cosmological term and bulk stress, Gen. Relativ. Gravit. **28**, 679–689 (1996)
29.42 A.I. Arbab, A. Beesham: Causal dissipative cosmology with variable G and Λ, Gen. Relativ. Gravit. **32**, 615–620 (2000)
29.43 A. Tawfik, M. Wahba, H. Mansour, T. Harko: Hubble parameter in QCD universe for finite bulk viscosity, Ann. Phys. **522**, 912–923 (2010)
29.44 F. Karsch, D. Kharzeev, K. Tuchin: Universal properties of bulk viscosity near the QCD phase transition, Phys. Lett. B **663**, 217–221 (2008)
29.45 M. Cheng, N.H. Christ, S. Datta, J. van der Heide, C. Jung, F. Karsch, O. Kaczmarek, E. Laermann, R.D. Mawhinney, C. Miao, P. Petreczky, K. Petrov, C. Schmidt, W. Soeldner, T. Umeda: The QCD equation of state with almost Physical quark masses, Phys. Rev. D **77**, 014511 (2008)
29.46 V.A. Belinski, I.M. Khalatnikov: Influence of viscosity on the character of cosmological evolution, Sov. Phys. JETP **42**, 205–210 (1976)
29.47 M. Heller: Singularities in Viscous Universes, Acta Cosmol. **7**, 7–15 (1978)
29.48 A. Woszczyna, W. Betkowski: Dynamics of viscous universe, Astrophys. Space Sci. **82**, 489–493 (1982)
29.49 N. Mostafapoor, Ø. Grøn: Bianchi type I universe models with nonlinear viscosity, Astrophys. Space Sci. **343**, 423–434 (2012)
29.50 Ø. Grøn: Expansion isotropization during the inflationary era, Phys. Rev. D **32**, 2522–2527 (1985)
29.51 R. Bali, P. Singh, J.P. Singh: Viscous Bianchi type I universe with stiff matter and decaying vacuum energy density, ISRN Math. Phys. 2012 (2012), 704612
29.52 G.P. Singh, A.Y. Kale: Anisotropic bulk viscous cosmological models with variable G and Λ, Int. J. Theor. Phys. **48**(4), 1177–1185 (2009)
29.53 R.K. Dubey, B.K. Singh, A. Mitra: An analysis of anisotropic cosmological model of universe with variable cosmological term, Int. J. Appl. Eng. Technol. 1(1), 7–14 (2011)
29.54 A. Pradhan, P. Pandey: Some Bianchi type I viscous fluid cosmological models with variable cosmological constant, Astrophys. Space Sci. **301**(1–4), 127–134 (2006)
29.55 A. Beesham, S.G. Ghosh, R.G. Lombard: Anisotropic viscous cosmology with variable G and Λ, Gen. Relativ. Gravit. **32**(3), 471–477 (2000)
29.56 G.P. Singh, S. Kumar: Bianchi type-I space-time with variable cosmological constant, Int J. Theor. Phys. **48**(8), 2401–2411 (2009)
29.57 L.D. Landau, E.M. Lifshitz: *Fluid Mechanics* (Addison-Wesley, Reading 1958)
29.58 I. Müller: Zum Paradoxon der Wärmeleitungstheorie, Z. Phys. A **198**(4), 329–344 (1967)
29.59 W. Israel: Nonstationary irreversible thermodynamics: A Causal relativistic theory, Ann. Phys. **100**, 310–331 (1976)
29.60 W. Israel, J. Stewart: Transient relativistic thermodynamics and kinetic theory, Ann. Phys. **118**, 341–372 (1979)
29.61 D. Pavón, D. Jou, J. Casas-Vázquez: On a covariant formulation of dissipative phenomena, Ann. Inst. Henri Poincaré (A) **36**(1), 79–88 (1982)
29.62 W. Hiscock, L. Lindblom: Stability and causality in dissipative relativistic fluids, Ann. Phys. **151**, 466–496 (1983)
29.63 V. Belinskii, E. Nikomarov, I. Khalatnikov: Investigation of the cosmological evolution of viscoelastic matter with causal thermodynamics, Sov. J. Exp. Theor. Phys. **50**, 213 (1979)
29.64 W. Zimdahl, J. Bafaluy, D. Jou: Cosmological particle production, causal thermodynamics, and inflationary expansion, Phys. Rev. D **61**, 083511 (2000)
29.65 D. Pavón, J. Bafaluy, D. Jou: Causal Friedmann–Robertson–Walker cosmology, Class. Quantum Gravity **8**, 347–360 (1991)
29.66 L. Chimento, A.S. Jakubi: Cosmological solutions of the Einstein equations with a causal viscous fluid, Class. Quantum Gravity **10**, 2047–2058 (1993)
29.67 L.P. Chimento, A.S. Jakubi: Dissipative cosmological solutions, Class. Quantum Gravity **14**, 1811–1820 (1997)
29.68 M. Zakari, D. Jou: Equations of state and transport equations in viscous cosmological models, Phys. Rev. D **48**(4), 1597–1601 (1993)
29.69 M. Mak, T. Harko: Full causal bulk-viscous cosmological models, J. Math. Phys. **39**, 5458 (1998)
29.70 A. Di Prisco, L. Herrera, J. Ibanez: Qualitative analysis of dissipative cosmologies, Phys. Rev. D **63**, 023501 (2001)
29.71 M. Szydlowski, O. Hrycyna: Dynamical dark energy models: Dynamical system approach, Gen. Relativ. Gravit. **38**, 121–135 (2006)
29.72 J. Belinchon: On the equation of state of a flat FRW model filled with a bulk viscous fluid (2005), arXiv:gr-qc/0412092
29.73 V. Folomeev, V. Gurovich: Viscous dark fluid, Phys. Lett. B **661**, 75–77 (2008)
29.74 A. Tawfik, H. Mansour, M. Wahba: Hubble parameter in bulk viscous cosmology, Proc. 12th Marcel Grossmann Meeting Gen. Relativ. (2012), arXiv: 0912.0115
29.75 J.C. Fabris, S. Goncalves, R. de Sa Ribeiro: Bulk viscosity driving the acceleration of the universe, Gen. Relativ. Gravit. **38**, 495–506 (2006)
29.76 B. Li, J.D. Barrow: Does bulk viscosity create a viable unified dark matter model?, Phys. Rev. D **79**, 103521 (2009)
29.77 W. Hipolito-Ricaldi, H. Velten, W. Zimdahl: The viscous dark fluid universe, Phys. Rev. D **82**, 063507 (2010)
29.78 O.F. Piattella, J.C. Fabris, W. Zimdahl: Bulk viscous cosmology with causal transport theory, J. Cosmol. Astropart. Phys. **1105**, 029 (2011)

29.79 W. Zimdahl, H.E.S. Velten, W.S. Hipólito-Ricaldi: Viscous dark fluid universe: A unified model of the dark sector?, Int. J. Mod. Phys. **3**, 312–323 (2011)
29.80 W. Zimdahl, D. Pavón, R. Maartens: Reheating and causal thermodynamics, Phys. Rev. D **55**, 4681–4688 (1997)
29.81 I. Prigogine, J. Geheniau, E. Gunzig, P. Nardone: Thermodynamics and cosmology, Gen. Relativ. Gravit. **21**, 767–776 (1989)
29.82 M.O. Calvao, J.A.S. Lima, I. Waga: On the thermodynamics of matter creation in cosmology, Phys. Lett. A **162**, 223–226 (1992)

30. Friedmann–Lemaître–Robertson–Walker Cosmology

David Wands

Presented is a discussion of homogeneous and isotropic cosmologies described by the Friedmann–Lemaître–Robertson–Walker (FLRW) metric. The cosmological models provide the framework within which astronomical observations of the Hubble expansion, cosmic microwave background radiation and primordial nucleosynthesis can be described. I present simple cosmological solutions of the Einstein equations in the case of vacuum spacetimes, radiation and dust, and discuss how an accelerated expansion (*inflation*) can solve some problems of the hot big bang model. In particular I discuss inhomogeneous perturbations about the FLRW background and how inflationary cosmology provides a model for the origin and evolution of structure in our Universe.

30.1 Motivation

The standard Hot Big Bang model of cosmology is based upon a simple class of spacetimes described by the Friedmann–Lemaître–Robertson–Walker (FLRW) metrics. While general relativity can, in principle, describe far more complicated spacetimes, symmetries are often used to reduce Einstein's field equations to something more manageable. Cosmologists appear to be fortunate to live in a universe that, on the largest observable scales, seems to be

1. *Spatially homogeneous*, which means that its properties are invariant under spatial translations (it looks the same at all positions at a given cosmic time).
2. *Isotropic*, which means that it is rotationally invariant (it looks the same in all directions).

30.1.1 Symmetries

The former doesn't actually imply the latter (you can have a spatially homogeneous universe that is expanding at unequal rates in different directions) and the latter does not imply the former (it just requires spherical symmetry about a given point). However, if a system is isotropic about all points, then it must be homogeneous too.

To be precise, we will assume that there exists a foliation of three-dimensional spatial hypersurfaces, each labeled by a cosmic time coordinate t, which are maximally symmetric, characterized by an intrinsic curvature scalar κ, which is the same at all spatial points at a given cosmic time.

For example, the two-dimensional surface of a sphere is a maximally-symmetric surface with a positive curvature proportional to the inverse square of the radius, $\kappa \propto L^{-2}$. The infinitesimal distance between two points on the sphere, labeled by angular coordinates (θ, ψ) and $(\theta + \mathrm{d}\theta, \psi + \mathrm{d}\psi)$, is given

$$\mathrm{d}s^2 = L^2 \left(\mathrm{d}\theta^2 + \sin^2\theta \, \mathrm{d}\psi^2\right) .$$

Mathematically this is readily extended to a three-dimensional sphere, with radius L and coordinates (χ, θ, ψ)

$$\mathrm{d}s^2 = \mathrm{d}\chi^2 + L^2 \sin^2(\chi/L) \left(\mathrm{d}\theta^2 + \sin^2\theta \, \mathrm{d}\psi^2\right) ,$$

or a three-dimensional hyperboloid (with negative curvature)

$$\mathrm{d}s^2 = \mathrm{d}\chi^2 + L^2 \sinh^2(\chi/L) \left(\mathrm{d}\theta^2 + \sin^2\theta \, \mathrm{d}\psi^2\right) .$$

On scales much smaller than the radius of curvature $\chi \ll L$, both cases reduce to flat three-dimensional space

$$\mathrm{d}s^2 = \mathrm{d}\chi^2 + \chi^2 \left(\mathrm{d}\theta^2 + \sin^2\theta \, \mathrm{d}\psi^2\right) .$$

We may write all three cases in unified form by introducing the angular diameter coordinate

$$r \equiv \begin{cases} L\sin(\chi/L) \\ \chi \\ L\sinh(\chi/L) \end{cases} , \tag{30.1}$$

such that

$$\mathrm{d}s^2 = \frac{\mathrm{d}r^2}{1-\kappa r^2} + r^2 \left(\mathrm{d}\theta^2 + \sin^2\theta \, \mathrm{d}\phi^2\right) ,$$

with $\kappa = +1/L^2$, 0 or $-1/L^2$ for spherical, flat or hyperbolic space, with positive, zero or negative spatial curvature, respectively.

These three-dimensional spaces have a fixed geometry, but our observed universe evolves with time. Thus we consider spatial foliations (three-dimensional slices of spacetime) with an overall scale factor that evolves with time, $a(t)$. The spacetime interval between neighboring events in the most general spatially homogeneous and isotropic spacetime is given by the FLRW metric [30.1]

$$\mathrm{d}s^2 = -c^2\,\mathrm{d}t^2 + a^2(t) \left(\frac{\mathrm{d}r^2}{1-\kappa r^2} + r^2\,\mathrm{d}\theta^2 + r^2 \sin^2\theta \, \mathrm{d}\phi^2\right) . \tag{30.2}$$

where $a(t)$ is the *scale factor* of the universe. The proper (physical) distance d, between any two observers *at rest* with respect to the homogeneous matter can be given as $\mathrm{d}(t) = a(t)\,\chi$, where χ is a fixed coordinate distance. The coordinate distance χ is also referred to as the *comoving distance* as it remains a constant for particles comoving with the cosmological expansion. The present value of the scale factor is taken to be $a_0 = 1$ and comoving distances then correspond to the physical distance at the present cosmic time.

If $\kappa > 0$ then the spatial geometry becomes analogous to geometry on the surface of a sphere. The angles of a triangle add up to more than 180° and the circumference of a circle is less than 2π times its radius. Note that the volume of space (at a fixed time t) is finite, and these models are sometimes referred to as *closed*.

If $\kappa < 0$ the space has a negative curvature (hyperbolic geometry). The angles of a triangle add up to less than 180° and the circumference of a circle is more than 2π times its radius. Like flat space, the volume of space at any fixed time t, may be infinite so these models are sometimes referred to as *open*. However, it is possible to construct flat or hyperbolic spaces with closed spatial topology [30.2].

The spatial curvature becomes significant over physical distances of order of the curvature scale $\mathrm{d}_\kappa = a/\sqrt{|\kappa|}$, or larger. Well within this scale the space still looks flat, just as the surface of the Earth may appear flat on scales less than $L_{\mathrm{earth}} = 6400\,\mathrm{km}$.

30.1.2 Cosmological Redshift

The original observational motivation to consider dynamical models of cosmology was the discovery of the redshift of light from distant galaxies, which is most simply interpreted as the expansion of our Universe.

Light rays follow a light-like or null path, i. e., the interval $\mathrm{d}s^2 = 0$. In the FLRW metric (30.2) this implies that for a photon traveling radially inwards towards $r = 0$, we must have

$$c^2\,\mathrm{d}t^2 = a^2(t)\,\frac{\mathrm{d}r^2}{1-\kappa r^2} . \tag{30.3}$$

We can show that this leads to the wavelength of light being stretched (or equivalently its frequency is *redshifted*) due to the expansion of the scale factor, as seen by observers comoving with the cosmic expansion.

Consider a light-ray emitted at some time t_e from a galaxy at $r = r_\mathrm{e}$. If it observed by us at time t_0, at

$r = 0$, then it must travel a coordinate distance

$$\chi_e = \int_0^{r_e} \frac{dr}{\sqrt{1-\kappa r^2}} = \int_{t_e}^{t_0} \frac{c\,dt}{a(t)} . \qquad (30.4)$$

Another light ray emitted from the same galaxy shortly afterwards at $t_e + \Delta t_e$ reaches us at $t_0 + \Delta t_0$. The coordinate distance is the same, so we also have

$$\chi_e = \int_{t_e+\Delta t_e}^{t_0+\Delta t_0} \frac{c\,dt}{a(t)} = \int_{t_e}^{t_0} \frac{c\,dt}{a(t)} , \qquad (30.5)$$

which implies

$$\int_{t_0}^{t_0+\Delta t_0} \frac{c\,dt}{a(t)} = \int_{t_e}^{t_e+\Delta t_e} \frac{c\,dt}{a(t)} . \qquad (30.6)$$

If we choose a small time interval Δt_e compared with the rate of expansion of the Universe then we can treat $a(t)$ as constant over this short interval. We then have

$$\frac{c\Delta t_0}{a(t_0)} = \frac{c\Delta t_e}{a(t_e)} . \qquad (30.7)$$

Therefore, light emitted with a frequency $\nu_e = 1/\Delta t_e$ is observed to have been redshifted when it is observed in an expanding universe

$$\frac{\nu_e}{\nu_0} = \frac{\Delta t_0}{\Delta t_e} = \frac{a(t_0)}{a(t_e)} . \qquad (30.8)$$

This redshift is seen by all observers who are at rest with respect to a homogeneous expansion, and appears as a Doppler shift of the frequency of light from sources receding from the observer. Distant galaxies are moving away from us because space itself is expanding.

We commonly refer to earlier times in terms of this redshift relative to the present (denoted by subscript 0)

$$1 + z_e \equiv \frac{\lambda_0}{\lambda_e} = \frac{\nu_e}{\nu_0} = \frac{a_0}{a_e} . \qquad (30.9)$$

For small distances, $t_e \simeq t_0$, we can expand

$$a(t_e) = a(t_0) - \left(\frac{da}{dt}\right)_0 (t_0 - t_e) + \dots . \qquad (30.10)$$

Hence

$$1 + z = \frac{a(t_0)}{a(t_e)} = 1 + H_0(t - t_e) + \dots , \qquad (30.11)$$

where the present expansion rate is given by the *Hubble constant*, $H_0 = (\dot{a}/a)_0$. For sufficiently small distances we recover Hubble's law that redshift is proportional to distance

$$z \simeq \frac{H_0\,d}{c} , \qquad (30.12)$$

where $d \simeq c(t_0 - t_e)$. What we refer to as the *Hubble constant* H_0 is just the present value of what is, in general, a time-dependent expansion rate,

$$H = \left(\frac{\dot{a}}{a}\right) . \qquad (30.13)$$

The linear Hubble law (30.12) is an approximate expression for small distances and for larger distances (higher redshifts) we need a more careful definition of distance in an expanding universe and in the presence of spatial curvature. One practical definition of distance is the luminosity distance d_L, defined in terms of the energy flux per unit area F, received from an object of absolute luminosity L

$$d_L = \sqrt{\frac{L}{4\pi F}} . \qquad (30.14)$$

In an FLRW metric (30.2) this gives

$$d_L = (1 + z)a_0 r(\chi) . \qquad (30.15)$$

Compared with the corresponding expression in flat Minkowski spacetime, there is an additional factor of $(1+z)$ due to the cosmological redshift which decreases the energy of photons observed and increases the observed interval between photons being emitted. $r(\chi)$ is the angular diameter distance (30.1) in curved space, where we can re-write (30.4) in terms of redshift as

$$\chi = \int_{t_e}^{t_0} \frac{c\,dt}{a(t)} = \int_0^z \frac{c\,dz}{a_0 H} . \qquad (30.16)$$

We can Taylor expand the Hubble expansion rate

$$\begin{aligned} H &\simeq H_0 - \dot{H}_0(t_0 - t) \\ &\simeq H_0[1 + (1 + q_0)z] , \end{aligned} \qquad (30.17)$$

where we define the dimensionless *deceleration parameter*

$$q_0 \equiv -\left(\frac{a\ddot{a}}{\dot{a}^2}\right)_0 . \qquad (30.18)$$

Thus observations of the luminosity distance (30.15) against redshift (30.16) can, in principle, measure not

only the present Hubble expansion, H_0, but also the deceleration, q_0.

30.1.3 Observational Cornerstones

Having introduced some of the key mathematical definitions, let us now briefly review some of the supporting evidence for spatially homogeneous and isotropic cosmologies.

Hubble Expansion

The most obvious test of the isotropy of our universe is to look at the position and motion of galaxies on the sky. They are not uniformly distributed, but they do appear to be *statistically* isotropic. Thus, the *average* number density of galaxies seems to be independent of direction, and indeed the types of galaxies found seem to be much the same in whatever direction we look. Nowadays not only the positions on the sky but also the redshifts of hundreds of thousands of galaxies are measured in systematic large-scale surveys [30.3, 4]. Combined with independent distance estimators this provides a strong confirmation of Hubble's expansion law (30.12).

For many years there was controversy about the correct value for the Hubble constant due to different methods for determining the distances to galaxies. A reliable determination of the Hubble constant was given by one of the cornerstone projects of the Hubble Space Telescope which used Cepheid variable stars in more distant galaxies to fix a value [30.5]

$$H_0 = 72 \pm 8\,\mathrm{km\,s^{-1}\,Mpc^{-1}}\,. \tag{30.19}$$

More recently effort has moved on to determine possible corrections to the linear Hubble law at high redshifts which could reveal spatial curvature and/or time-dependence of the Hubble rate as the cosmic expansion slowed down, or speeded up. Type IA supernovae are exploding stars which are visible out to redshifts of order 1. They are thought to originate from stars close to the Chandrasekhar mass, approximately 1.4 times the mass of our Sun, and hence are all of similar intrinsic luminosity. Observations of local supernovae have been used to calibrate the intrinsic brightness against the rate of decay of the supernova light-curve and further reduce the scatter in the intrinsic luminosity. Surveys of distant type IA supernovae have been used to show that the expansion of the universe is accelerating [30.6], confounding expectations of a universe filled with ordinary matter and radiation, with no evidence of spatial curvature.

We now have so much data about the angular positions (on the sky) and redshifts of galaxies that even without independent estimates of their distance, galaxy redshift surveys are used to test cosmological theories and constrain model parameters [30.3, 7].

The Cosmic Microwave Background

Just because our universe is expanding does not mean there had to be a big bang. For many years the steady state theory offered an alternative explanation for the expanding universe [30.8]. It proposed that matter was continually being created, so the overall state of the universe was uniform in time as well as space.

The primary observational grounds for believing that the universe emerged from a hot big bang is the presence of the cosmic microwave background (CMB). Today these microwave photons are part of a thermal spectrum of radiation corresponding to a temperature of 2.73 K [30.9]. However, we know that the frequency of photons is redshifted by the expansion of the universe, (30.9), and so the CMB temperature is inversely proportional to the size of the universe,

$$\frac{T}{T_0} = \frac{a_0}{a}\,. \tag{30.20}$$

The CMB was discovered by *Penzias* and *Wilson* in 1965 [30.10] as a uniform background radiation. Apart from a small dipole component (produced by our motion relative to the Hubble flow), the primordial radiation is observed to be isotropic to about 1 part in 10^5. This is consistent with the idea that the early universe was nearly homogeneous with only small perturbations. Over time these small perturbations can grow, due their own gravitational attraction, to produce the distribution of galaxies that we observe today.

In 1992 the first evidence of primordial anisotropies in the microwave background sky were reported by the Cosmic Background Explorer (COBE) satellite [30.11]. This has led to a new era in observational cosmology. A large number of balloon, ground-based and satellite experiments have confirmed COBE's measurements and extended them to precise measurements of the temperature and polarization anisotropies in the CMB across a range of angular scales. These observations have given us a snapshot of the Universe was when it was only a few hundred thousand years old. The specific features of this map give us a stringent test of different cosmological models and parameters [30.12]. The general picture of an isotropic universe very close to thermal equilibrium remains an

essential feature of current models of the very early universe.

Primordial Nucleosynthesis

About the same time that the microwave background was first discovered in the 1960s, theorists were beginning to take seriously another prediction of the hot big bang model. If temperatures really were so high in the early universe then there must have been an epoch at which nuclear reactions were in thermal equilibrium. A hot plasma of electrons and protons, photons, and neutrons and neutrinos, all interacting and cooling as the universe expanded makes specific predictions for the primordial abundances of the elements.

The early universe turns out to be successful at producing nuclei of the lightest elements: helium and lithium, as well as hydrogen, but heavier elements are produced later in stars. By studying the abundance of the light elements and their different isotopes, and plotting this against the abundance of heavier elements one can extrapolate back to a primordial value (where there were no heavy elements). The relative abundances of the light elements are sensitive to the cosmology only one second after the big bang [30.13]. Changing the rate of cosmological expansion or the relative number of photons to protons, say, affects these yields. From the observed abundances of light elements we can see that the primordial composition of distant galaxies appears to have been much the same as our own galaxy. This provides evidence that the early universe, when these elements were made, was indeed homogeneous.

30.2 Dynamical Equations and Simple Solutions

Einstein's field equations tell us how the spacetime curvature is determined by the energy-momentum of matter. In a homogeneous and isotropic universe (30.2) the symmetry of spacetime leaves us with only two independent equations. One is the evolution equation which determines the acceleration of the scale factor

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + 3\frac{p}{c^2}\right) + \frac{\Lambda}{3}\,, \tag{30.21}$$

where ρ is the matter density and p the pressure (which is required to be isotropic), and we have allowed for a cosmological constant Λ in the Einstein equations. The other equation is the Friedmann constraint equation, which shows how the density and spatial curvature determine the expansion rate at any time

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{\kappa c^2}{a^2} + \frac{\Lambda}{3}\,. \tag{30.22}$$

The Friedmann constraint may also be derived as a first integral of the evolution equation (30.21) using the matter continuity equation

$$\dot{\rho} = -3H\left(\rho + \frac{p}{c^2}\right)\,. \tag{30.23}$$

Through solving these equations of motion for the scale factor in a homogeneous universe we can understand how our universe may have evolved from an initial big bang to the present.

30.2.1 True Vacuum

In the absence of any energy density (a true vacuum spacetime with $\rho = 0$, $P = 0$ and $\Lambda = 0$) we find two solutions of the Friedmann equation (30.22). One has no spatial curvature $\kappa = 0$ and hence $\dot{a} = 0$ and, as we might expect, we find static Minkowski spacetime with

$$\begin{aligned} ds^2 &= -c^2\,dt^2 + dr^2 \\ &\quad + r^2\left(d\theta^2 + \sin^2\theta\,d\phi^2\right)\,. \end{aligned} \tag{30.24}$$

However, the second solution is obtained with $\kappa = -1/L^2$ and $\dot{a} = c/L$, hence

$$\begin{aligned} ds^2 &= -c^2\,d\tilde{t}^2 + \left(\frac{c\tilde{t}}{L}\right)^2 \\ &\quad \times \left[d\chi^2 + L^2\sinh^2(\chi/L)\right. \\ &\quad \left.\times\left(d\theta^2 + \sin^2\theta\,d\phi^2\right)\right]\,. \end{aligned} \tag{30.25}$$

This is the Milne universe, where the spacetime curvature is zero, but spatial slices at a given time $\tilde{t}$ have negative spatial curvature. Comoving observers, following geodesics orthogonal to these spatial slices, see a uniform expansion. It can be shown that this is, in fact, a coordinate transform of Minkowski spacetime (30.24) where

$$\begin{aligned} ct &= c\tilde{t}\cosh(\chi/L)\,, \\ r &= c\tilde{t}\sinh(\chi/L)\,. \end{aligned} \tag{30.26}$$

The Milne coordinates $(\tilde{t}, \chi)$, describe only the interior of the light-cone $r^2 < c^2t^2$ in Minkowski spacetime. In this case the apparent Big Bang singularity in the Milne universe where the scale factor vanishes at $\tilde{t} = 0$ is, in fact, only a coordinate singularity (at $r = ct$ in the Minkowski chart (30.24)) and the density and pressure remain zero throughout.

30.2.2 Radiation

The Hot Big Bang cosmology describes a universe filled by a fluid of relativistic particles, such as photons. An isotropic distribution in three-dimensional space exerts a pressure

$$p = \frac{1}{3}\rho c^2 . \tag{30.27}$$

The factor of $1/3$ simply reflects the average component of each particle's total momentum along each spatial direction. The continuity equation (30.23) then gives

$$\dot{\rho} = -4\frac{\dot{a}}{a}\rho , \tag{30.28}$$

which can be integrated to yield

$$\rho = \frac{\Gamma}{a^4} , \tag{30.29}$$

where Γ is a constant.

Another way to derive this relation between density and scale factor is to realize that, if the total photon number is conserved, the photon number density must decrease as $n \propto a^{-3}$, while the energy of each individual photon decreases as $h\nu \propto a^{-1}$ due to being redshifted as its wavelength is stretched by the cosmological expansion. Thus we have

$$\rho c^2 \propto n\, h\nu \propto a^{-4} . \tag{30.30}$$

Substituting this into the Friedmann constraint equation (30.22), taking $\Lambda = 0$, and multiplying through by a^4 gives

$$\left[\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(a^2\right)\right]^2 = \frac{8\pi G\Gamma}{3} - \kappa c^2 a^2 . \tag{30.31}$$

This can be integrated to give

$$a^2 = \left[\sqrt{\frac{32\pi G\Gamma}{3}} - \kappa c^2(t - t_*)\right](t - t_*) . \tag{30.32}$$

The qualitative evolution of $a(t)$ for $t > t_*$ depends solely on the sign of κ:

1. If $\kappa < 0$ the scale factor, expanding from zero at $t = t_*$, approaches an asymptotically constant speed $\dot{a} \to \sqrt{-\kappa}c$ as $a \to \infty$.
2. If $\kappa > 0$ then the scale factor only reaches a maximum size $a_{max} = \sqrt{8\pi G\Gamma/3\kappa c^2}$ before recollapsing back to zero size at $t - t_* = (\sqrt{32\pi G\Gamma/3})/\kappa c^2$.
3. If $\kappa = 0$ we have
$$a \propto (t - t_*)^{1/2} . \tag{30.33}$$
The scale factor grows to an infinite size in an infinite time, but the limiting speed $\dot{a} \to 0$ as $a \to \infty$.

Note that models with $\kappa \neq 0$ remain close to the $\kappa = 0$ solution at early times ($|\kappa c^2(t - t_*)| \ll \sqrt{32\pi G\Gamma/3}$) but then diverge from this behavior at late times.

All these solutions have a Big Bang singularity where the scale factor a vanishes at $t \to t_*$. Unlike the Milne universe, this is curvature singularity where the density and pressure diverge. In the spherical universe, $\kappa > 0$, there is also a big crunch singularity when the scale factor recollapses to zero size.

30.2.3 Dust

The energy density of cold (nonrelativistic) matter ($v \ll c$) is dominated by its rest-mass energy and the pressure it exerts is negligible ($p \ll \rho c^2$). Such a pressureless fluid is referred to as *dust*.

The continuity equation (30.23) gives

$$\dot{\rho} = -3\frac{\dot{a}}{a}\rho , \tag{30.34}$$

and thus

$$\rho = \frac{\Gamma}{a^3} . \tag{30.35}$$

As in the case of radiation this corresponds to the number of particles being conserved $n \propto a^{-3}$, but unlike radiation the energy of nonrelativistic particles is not affected by the redshift, so

$$\rho = n m_0 \propto a^{-3} . \tag{30.36}$$

Substituting this into the constraint equation (30.22), with $\Lambda = 0$, gives

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\Gamma}{3a^3} - \frac{\kappa c^2}{a^2} . \tag{30.37}$$

The simplest case to integrate is when $\kappa = 0$. Then we have

$$\int a^{1/2}\mathrm{d}a = \sqrt{\frac{8\pi G\Gamma}{3}}\int \mathrm{d}t , \tag{30.38}$$

and so

$$a=\left[\frac{3}{2}\sqrt{\frac{8\pi G\Gamma}{3}}(t-t_*)\right]^{2/3} . \tag{30.39}$$

Analytic solutions for dust cosmologies with spatial curvature $\kappa \neq 0$ can be given in parametric form [30.14]. They obey the same qualitative picture as seen for the radiation solutions for $\kappa > 0$ and $\kappa < 0$.

30.2.4 Barotropic Fluids

Any fluid obeying an equation of state $p = w\rho c^2$ where the equation of state w is a function of the density is referred to as a barotropic fluid. These are very useful in cosmology, partly because they include the two important cases given above, but also because whenever w is a constant we can integrate the continuity equation to find

$$\rho = \frac{\Gamma}{a^{3(1+w)}} . \tag{30.40}$$

This in turn allows us to integrate the Friedmann constraint equation (30.22) when $\Lambda = 0$ and $\kappa = 0$, to give $a \propto (t-t_*)^{2/3(1+w)}$.

Note that for $\Lambda = 0$ and $\kappa = 0$ models with barotropic equations of state we always have a power-law expansion of the scale factor with respect to the time t. Moreover, we find the density decreases, $\rho \propto (t-t_*)^{-2}$, and the Hubble rate can always be written as

$$H=\frac{\dot{a}}{a}=\frac{2}{3(1+w)(t-t_*)} . \tag{30.41}$$

The age of the universe is inversely proportional to the expansion rate.

The acceleration of the scale factor, $\ddot{a}$ in (30.21), is always negative for $p > -\rho c^2/3$, i. e., $w > -1/3$. To explain the acceleration of the universe indicated by high-redshift supernovae surveys, within FLRW models in general relativity with $\Lambda = 0$, requires an exotic equation of state with $w < -1/3$.

30.2.5 False Vacuum

The simplest explanation for the apparent acceleration of the universe today could be provided by a nonzero energy density of empty space. If the vacuum has a nonzero energy density, $\rho = \rho_V =$ constant, then it would remain undiluted by cosmological expansion, $\dot{\rho} = 0$, and hence from (30.23) the vacuum must have $p = -\rho c^2$. We have, in effect, a barotropic fluid with equation of state $w = -1$.

We have already considered true vacuum solutions where $\rho_V = 0$, but it is possible to have $\rho_V \neq 0$ and this is referred to as the false vacuum, to distinguish it from the true vacuum where $\rho = 0$. In most of physics, the absolute value of the vacuum energy is irrelevant, and all that matters is the change in energy due to different physical processes. However, in general relativity it is the total energy density, including the vacuum energy that appears in the field equations. Such a term is identical to a cosmological constant, Λ, in the Einstein field equations. In particular, the Friedmann equation (30.22) with a nonzero vacuum energy gives

$$H^2=\frac{\Lambda}{3}-\frac{\kappa c^2}{a^2} , \tag{30.42}$$

where $\Lambda = 8\pi G\rho_V$.

This can be integrated for $\Lambda > 0$ to give

$$a(t)=\begin{cases} a_*\cosh\left[H_{\mathrm{dS}}(t-t_*)\right] & \text{for } \kappa=+1 \\ a_0\exp\left[H_{\mathrm{dS}}(t-t_0)\right] & \text{for } \kappa=0 \\ a_*\sinh\left[H_{\mathrm{dS}}(t-t_*)\right] & \text{for } \kappa=-1 \end{cases} , \tag{30.43}$$

where $H_{\mathrm{dS}} = \sqrt{\Lambda/3}$ and $a_*^2 = 3|\kappa|c^2/\Lambda$.

This is the de Sitter spacetime first derived by de Sitter in 1917. The intrinsic spacetime curvature is constant throughout time as well as space. The apparent singularity when $a = 0$ at $t = t_*$ for $\kappa < 0$ is, in fact, a coordinate singularity analogous to the coordinate singularity in the Milne universe (30.25).

As $t \to \infty$ all the solutions approach exponential expansion. Thus all models approach the spatially flat $\kappa = 0$ solution at late times. In particular, the acceleration of the scale factor is always positive, $\ddot{a} > 0$, in contrast with all the models we have considered up to this point.

For $\Lambda = 0$ and $\kappa \leq 0$, we recover the true vacuum solutions discussed earlier, while $\Lambda < 0$ gives anti-de Sitter solutions which are not thought to describe our four-dimensional cosmology, but play a central role in many theoretical studies of supergravity theories and holography [30.15].

30.3 The Density Parameter Ω

Two long-standing challenges in observational cosmology are to determine the value of the cosmological constant Λ and the spatial curvature of our Universe κ in the FLRW models. One way to quantify their effect is to measure the expansion rate H and compare it with the observed matter density.

If the universe were spatially flat ($\kappa = 0$) with vanishing cosmological constant ($\Lambda = 0$), the Friedmann equation (30.22) would require a *critical density*

$$\rho_c \equiv \frac{3H^2}{8\pi G} . \quad (30.44)$$

For the present-day value of the Hubble constant this implies

$$\rho_{c0} = 1.9 \times 10^{-29} h^2 \text{g cm}^{-3} . \quad (30.45)$$

where the dimensionless Hubble constant $h = H_0/(100\,\text{km s}^{-1}\,\text{Mpc}^{-1})$.

The actual density of matter relative to this critical density is denoted by

$$\Omega \equiv \frac{8\pi G\rho}{3H^2} . \quad (30.46)$$

It has become common to define an analogous parameter for the cosmological constant (or equivalently a false vacuum energy density)

$$\Omega_\Lambda \equiv \frac{\Lambda}{3H^2} . \quad (30.47)$$

The sum of all the different energy densities ρ_i and the cosmological constant is then given by

$$\Omega_{\text{tot}} = \Omega_\Lambda + \sum_i \Omega_i . \quad (30.48)$$

The value of the dimensionless parameter Ω shows the relative contribution of the energy density to the Hubble expansion in the constraint equation, (30.22):

1. $\Omega_{\text{tot}} = 1$: The spatial metric is flat ($\kappa = 0$) so the expansion is solely driven by the matter density and/or the cosmological constant.
2. $\Omega_{\text{tot}} < 1$: The spatial metric is hyperbolic ($\kappa < 0$) so the expansion is partly due to the curvature.
3. $\Omega_{\text{tot}} > 1$: The spatial metric is spherical ($\kappa > 0$) so there is more than enough matter and/or cosmological constant to *close the universe*. The curvature leads to a slower expansion than would be inferred solely from the energy density and cosmological constant.

Although spatial curvature and the cosmological constant both affect the Hubble expansion through the Friedmann equation (30.22), the curvature also determines the geometry of space. The cosmic microwave background temperature anisotropies play a key role in determining this spatial curvature in our universe today. These anisotropies have a characteristic scale corresponding to the Hubble length, cH^{-1}, at last-scattering of the CMB photons, at a redshift $z_{\text{lss}} \simeq 1100$. This would correspond to an angular scale of approximately $1°$ in a spatially flat universe, but would coincide with a larger angular scale in a spherical geometry, $\kappa > 0$, or a smaller angular scale in a hyperbolic geometry. The fact that the observed angular scale closely matches the theoretical prediction in a flat geometry can be used to determine [30.16]

$$\Omega_{\text{tot}} = 1.04 \pm 0.04 . \quad (30.49)$$

The energy density of the CMB photons themselves can be calculated precisely from the observed black-body spectrum and temperature of the microwave background radiation

$$\rho_{\gamma 0} = 4.8 \times 10^{-34} \text{g cm}^{-3} , \quad (30.50)$$

which corresponds to

$$\Omega_{\gamma 0} = 2.6 \times 10^{-5} h^{-2} . \quad (30.51)$$

Thus radiation makes a very small contribution to the expansion rate in our present universe.

The present density of nonrelativistic matter in the universe is significantly larger than that of radiation. The density of luminous matter seen in galaxies, stars, and gas, principally baryonic matter (atomic nuclei) like our Sun

$$\Omega_{B0} \simeq 0.05 , \quad (30.52)$$

is consistent with that required by theoretical models of primordial nucleosynthesis [30.13]. However dynamical studies of galaxies and clusters of galaxies suggest that the real mass of matter in galaxies is much larger than this, requiring some form of nonbaryonic, *dark matter*.

If we assume a spatially flat geometry then an estimate of the total matter density in the universe today comes from comparing the observed angular power

spectrum of CMB anisotropies to theoretical predictions using different model parameters. Data from the WMAP satellite, for example, compared against theoretical predictions for a universe dominated by matter and/or a cosmological constant at the present time gives [30.16]

$$\Omega_{m0} = (0.136 \pm 0.005)h^{-2}\,, \tag{30.53}$$

$$\Omega_{\Lambda 0} = 0.72 \pm 0.03\,. \tag{30.54}$$

This is the basis of the standard cosmological model at present, commonly referred to as ΛCDM. Combining CMB data with additional data including type IA supernovae and galaxy redshift surveys gives even tighter constraints on parameters within a ΛCDM cosmology.

Remembering that the density of a relativistic fluid like photons decreases as a^{-4} as the universe expands (30.29) while nonrelativistic matter decreases as a^{-3} (30.35), we deduce that the two densities must have been equal at some time in the past t_{eq}, where

$$\begin{aligned} 1 + z_{\text{eq}} &\equiv \frac{a_0}{a_{\text{eq}}} = \frac{\Omega_{m0}}{\Omega_{\gamma 0}} = 3.9 \times 10^4 \Omega_{m0} h^2 \\ &\simeq 5300\,. \end{aligned} \tag{30.55}$$

At earlier times relativistic rather than nonrelativistic matter would have dominated the energy density of the universe. This is the basis of the Hot Big Bang cosmology.

On the other hand, a cosmological constant remains undiluted while the matter and radiation densities decrease as the universe expands. Therefore, Λ is apparently unimportant in the very early universe, but comes to play a dominant role in the present expansion. The fact that this occurs at a very recent cosmological epoch is known as the coincidence problem in modern cosmology. In the future, a cosmological constant will increasingly dominate the Hubble expansion and matter density will have a negligible effect.

30.3.1 The Flatness Problem

If Ω is not exactly equal to unity then why should it still be as close as 1.1 or 1.01 today? If the spatial curvature had been comparable to the matter density in the very early universe it would then have dominated the expansion within a few Hubble times. A spherical model would recollapse, while a hyperbolic model would expand so that the matter density would rapidly become negligible.

The Friedmann constraint equation (30.22) can be written as

$$\Omega = 1 + \frac{\kappa c^2}{\dot{a}^2}\,. \tag{30.56}$$

We see that Ω is driven away from 1 whenever $\dot{a}^2$ decreases.

The evolution equation (30.21) shows that $\dot{a}$ decreases as the universe expands with either radiation ($p = \rho c^2/3$) or dust domination ($p = 0$), so $|\Omega - 1|$ must grow with time in the early universe

$$\begin{aligned} \Omega - 1 &\propto a^2 \text{ for radiation dominated}\,, && (30.57) \\ &\propto a \text{ for dust dominated}\,. && (30.58) \end{aligned}$$

If we use the matter dominated solution for $T < T_{\text{eq}}$ and radiation dominated solutions for $T > T_{\text{eq}}$, we see that if $|1 - \Omega_0|$ is 0.01 today, then $|1 - \Omega|$ would have to be $0.01 \times (T_0/T_{\text{eq}}) \times (T_{\text{eq}}/10^{10}\text{K})^2 \approx 10^{-16}$ at the time of nucleosynthesis.

30.3.2 Inflation

False vacuum cosmologies, (30.43), with $\rho_V > 0$ approach the spatially flat solution at late times. Unlike radiation or dust solutions, these false vacuum solutions correspond to an accelerated expansion

$$\ddot{a} > 0\,. \tag{30.59}$$

This implies that $\dot{a} = aH$ is then an increasing function of time and $\Omega - 1 \to 0$ in (30.56).

The acceleration of the scale factor (30.59) can be taken as a definition of *inflation* in an FLRW cosmology. The quantity $c/\dot{a} = cH^{-1}/a$ is a comoving measure of the Hubble length, cH^{-1}. Thus we see that the comoving Hubble length decreases during inflation.

We see from the evolution equation (30.21) that inflation (30.59) occurs when

$$p < \frac{\Lambda c^2}{4\pi G} - \frac{1}{3}\rho c^2\,. \tag{30.60}$$

The present value of the cosmological constant would be negligible when compared with the energy density of matter in the very early universe, and hence in practice we require

$$p < -\frac{1}{3}\rho c^2 \tag{30.61}$$

for inflation.

A false vacuum energy density, $p = -\rho c^2$, provides one possible source of inflation. Taking $\dot{a} \propto a$ in a false vacuum dominated cosmology, we have

$$\Omega - 1 \propto a^{-2} . \qquad (30.62)$$

As long as inflation lasts long enough there is no problem in driving Ω arbitrarily close to one in the very early universe. However, a very early era of accelerated expansion must give way to the standard radiation and dust dominated eras, so the false vacuum energy must decay into ordinary matter, returning the universe to thermal equilibrium before the epoch of primordial nucleosynthesis, a process known as *reheating*.

30.4 Cosmological Horizons

The Hot Big Bang model provides a good description of many aspects of the observed universe, based on established models of particle interactions in a homogeneous and isotropic universe obeying the evolution equations of Einstein's general relativity. However, at the same time there are a number of unanswered questions.

Just why is our universe so homogeneous? For instance, the microwave background is sensitive to perturbations in the metric at the time of last scattering of the CMB photons. Perturbations in the gravitational potential produce temperature anisotropies, yet the microwave sky is uniform to about one part in a hundred thousand. On the other hand, the universe cannot be completely homogeneous or there would be no structures, such as galaxies and stars, today. Gravity tends to make matter clump together so should amplify any initial inhomogeneities. Small density perturbations in the dust dominated era grow proportional to the scale factor, so if the density perturbation $\delta\rho/\rho$ is only of order unity today, as we observe on scales of about $8h^{-1}$ Mpc, these must have grown from an initial density perturbation $\approx 10^{-4}$ at t_{eq}, see (30.55). Why such small initial inhomogeneities, and where did they come from?

In order to avoid producing large metric perturbations on small scales (which would collapse to form black holes) and to avoid large inhomogeneities on large scales (which would be seen on the CMB sky), an almost scale invariant distribution of density perturbations is required in order to produce structure across a wide range of scales. A *Harrison–Zel'dovich spectrum* corresponds to an exactly scale-invariant initial gravitational potential, $\phi \approx GL^2\delta\rho/c^2 \approx 10^{-4}$, on all scales L.

30.4.1 Particle Horizon

This smoothness problem becomes even worse if we think about the size of causally connected regions in the early universe. Starting from the Big Bang at $t_* = 0$ a light ray travels a coordinate distance equal to $\int_0^t c\,\mathrm{d}t'/a(t')$ which at time t, in the radiation dominated era (30.33), corresponds to a physical distance

$$\mathrm{d}(t) = a(t) \int_0^t \frac{c\,\mathrm{d}t'}{a(t')} = 2ct = cH^{-1} . \qquad (30.63)$$

This is called the particle horizon distance. At the time of nucleosynthesis, say, this corresponds to a physical distance of about 10^{12} cm. This region, stretched by the Hubble expansion up to the present, corresponds to about 10^{21} cm, which is only 1 kpc, less than the size of our galaxy.

Assuming that nothing travels faster than the speed of light there would seem to be no way that causally disconnected regions could have established homogeneity by that time, and yet there is no evidence for different primordial nuclear abundances in different parts of our galaxy, or even different galaxies. Causally disconnected regions could not know of one another's existence much less establish homogeneity and thermal equilibrium in the standard Hot Big Bang model. Similarly, if we estimate the size of causally connected regions on the microwave background sky at the time of decoupling, they are only about 1° across. The initial conditions must have been set up so that the universe was not only approximately homogeneous over causally disconnected regions, but also close to thermal equilibrium and contained a nearly Harrison–Zel'dovich spectrum of small but nonzero density perturbations.

30.4.2 Inflating Horizons

In calculating the particle horizon size during the early radiation dominated era we implicitly assumed that there was no significant contribution from any preceding era. Suppose there is an earlier nonradiation

dominated era, for $t < t_r$, then we can split the integral in (30.63) to give

$$\mathrm{d}(t) = a(t)\left[\int_{t_r}^{t} \frac{c\,\mathrm{d}t'}{a(t')} + \int_{t_i}^{t_r} \frac{c\,\mathrm{d}t'}{a(t')}\right]$$
$$= 2c(t-t_f) + \left(\frac{t}{t_r}\right)^{1/2} d_r\,. \qquad (30.64)$$

The second term can be neglected for conventional (noninflationary) evolution if $t_r \ll t$. For instance, for power-law evolution $a \propto (t-t_i)^n$, we have $d_r = c(t_r - t_i)/(1-n) < (n/(1-n))cH_r^{-1}$ for $n < 1$.

However, after a period of inflation this may no longer be the case. We can write

$$d_r = a_r \int_{t_i}^{t_r} \frac{c\,\mathrm{d}t'}{a(t')} = a_r \int_{a_i}^{a_r} \frac{c}{\dot{a}}\frac{\mathrm{d}a}{a}\,. \qquad (30.65)$$

During inflation $\dot{a}$ increases so $\dot{a}_r > \dot{a}$ at earlier times and hence

$$d_r > a_r \int_{a_i}^{a_r} \frac{c}{\dot{a}_r}\frac{\mathrm{d}a}{a} = cH_r^{-1}\ln\left(\frac{a_r}{a_i}\right). \qquad (30.66)$$

Thus in the limit $a_i \to 0$ the distance d_r is divergent and the particle horizon is undefined. Put another way, as the duration of inflation, represented here by the number of e-folds, $N_{\mathrm{inf}} = \ln(a_r/a_i)$ becomes arbitrarily large, so the physical scale of causally connected regions becomes arbitrarily large.

Thus it becomes possible to study causal models for the origin of large-scale structure across a wide range of scales in our universe if we allow for an inflationary epoch in the very early universe, before primordial nucleosynthesis.

30.5 Inhomogeneous Perturbations

While the FLRW metric provides an idealized description of a perfectly homogeneous and isotropic spacetime, our observed Universe has localized fluctuations in density and temperature. These may be described as inhomogeneous perturbations about a spatially homogeneous background, so that the local mass density, for example, is given by

$$\rho(t,x^i) = \bar{\rho}(t) + \delta\rho(t,x^i)\,. \qquad (30.67)$$

On large scales (greater than about 10 Mpc in the present Universe) or at early times, the local density fluctuations are small and their evolution may be described by linearized equations of motion (keeping only terms to first order in the perturbations).

Breaking the symmetry of the FLRW background cosmology breaks some important simplifications and re-introduces some complexity. In particular, quantities such as the density or pressure perturbation at a given spacetime point become gauge-dependent. The FLRW background has a preferred choice of cosmic time, corresponding to homogeneous spatial hypersurfaces at time t, but in the presence of inhomogeneities there is no unique choice of spatial hypersurfaces, and we can redefine our time coordinate

$$t \to t + \delta t(t,x^i)\,. \qquad (30.68)$$

This leads to a change in the local density perturbation due to the change in the split between background density and perturbed density. At first order we find

$$\delta\rho \to \delta\rho - \dot{\bar{\rho}}\delta t\,. \qquad (30.69)$$

Gauge-invariant perturbations can be defined by identifying perturbations on physically defined hypersurfaces [30.17], but there is no unique set of gauge-invariant perturbations, such as the gauge-invariant density perturbation, and different authors may use different gauge invariant quantities in different situations.

Inhomogeneities in the matter and metric can be decomposed into scalar, vector, and tensor modes defined with respect to the homogeneous background spatial metric [30.18]. Vector modes are *transverse* (i. e., divergence-free, $\nabla_i v^i = 0$), while tensor modes are *transverse and tracefree*. Scalars can represent scalar quantities, such as density or pressure, or tensorial quantities that can be constructed from spatial derivatives of scalars. For example, the three-velocity of a fluid can be decomposed into scalar and vector parts,

$$v^i = \nabla^i v^{(S)} + v^{(V)i}\,,$$

where the potential $v^{(S)}$ determines the scalar part and the vorticity $v^{(V)i}$ is transverse.

The linearized field equations can be split into scalar, vector, and tensor parts, and thus the scalar,

vector, and tensor perturbations evolve independently. One can study the evolution of the linear density perturbation coupled to scalar metric perturbations and the scalar velocity potential $v^{(S)}$, independently of the vector or tensor metric perturbations, or the transverse velocity $v^{(V)i}$. Scalar, vector, and tensor perturbations can be further decomposed into Fourier modes (eigenmodes of the spatial Laplacian) with wavenumber k.

30.5.1 Density Waves

Density waves with different comoving wavevectors $\boldsymbol{k}$ evolve independently in this linear regime in an FLRW background. The density contrast $\delta \equiv \delta\rho/\rho$ for adiabatic density perturbations in a comoving-orthogonal gauge (i. e., on constant time hypersurfaces orthogonal to observers comoving with the cosmic fluid) obeys the wave equation [30.19]

$$\ddot{\delta} + \left(2 - 6w + 3c_s^2\right) H\dot{\delta} + \left[\frac{c_s^2 k^2}{a^2} - \frac{3H^2}{2}\left(1 + 8w - 3w^2 - 6c_s^2\right)\right]\delta = 0\,, \tag{30.70}$$

where $w = p/\rho$ is the equation of state, $c_s^2 = \dot{p}/\dot{\rho}$ is the adiabatic sound speed, and $k = |\boldsymbol{k}|$ is the comoving wavenumber of the perturbation, corresponding to a physical wavelength $\lambda = 2\pi a/k$.

Spatial pressure gradients, $\nabla p = c_s^2 \nabla\rho$, drive oscillations on small scales, while the cosmological expansion leads to the Hubble damping term. The competition between these two effects divides the behavior of density waves into two regimes

- *Small-scale* underdamped oscillations for $c_s^2 k^2/a^2 > H^2$, whose amplitude decays $\propto 1/a$.
- *Large-scale* overdamped perturbations *frozen in* for $c_s^2 k^2/a^2 < H^2$.

In the early Hot Big Bang the sound speed for the hot relativistic plasma is $c_s^2 = c^2/3$. All comoving modes in the radiation dominated era, with $a \propto t^{1/2}$ and $aH \propto t^{-1/2}$ start frozen in, in the large-scale regime, and only oscillate after they *enter the horizon* when $c^2k^2/3a^2 = H^2$, corresponding to $\lambda \approx ct$.

Observations of the spectrum of temperature and polarization anisotropies in the CMB data, from the WMAP satellite for example, show a pattern of coherent oscillations, whose phase at the last scattering of the CMB photons depends upon the wavelength and hence the time each mode entered the horizon. This provides strong evidence for an initial almost scale-invariant spectrum of adiabatic density perturbations, existing on super-horizon scales before the time of last-scattering. The standard hot big bang model offers no explanation for how these primordial density perturbations could arise without apparently breaking causality.

In the dust-dominated era which follows the radiation-dominated era, with $w = 0$ and hence $a \propto t^{2/3}$, where nonrelativistic matter dominates the energy density and the sound speed drops close to zero, linear perturbations can grow relative to the background density on all scales with $\delta \propto a$. This is thought to be the origin of all the structure observed in galaxy redshift surveys in the present universe, having grown from small density perturbations at the start of the dust-dominated era.

The growth of structure, observed at different cosmological redshifts, provides a test of both the primordial distribution of density perturbations and also the relative energy densities controlling the dynamics at later cosmic times, including radiation, neutrinos, baryonic and cold or hot dark matter, and dark energy or modified gravity today.

30.5.2 Inflation and the Origin of Structure

Inflation offers a mechanism to generate the large scale structure observed in our Universe today from microscopic fluctuations in the very early universe. Inflation is a period of accelerated expansion in the early universe, such that $aH = \dot{a}$ grows with time. Hence wave modes that originate as oscillations on sub-Hubble scales with $c_s k/a > H$, are stretched up to super-Hubble scales with $c_s k/a < H$. In particular, inhomogeneous perturbations can originate from quantum fluctuations of free fields during a period of inflation. Assuming only that there exist zero-point vacuum fluctuations on small scales, $ck/a \ll H$ (for massless fields with $c_s = c$) this leads to field fluctuations proportional to the Hubble scale when modes *leave the horizon* with $ck/a = H$. These fluctuations then enter the overdamped regime, becoming squeezed in phase space (since the decaying mode is rapidly damped) and can be treated as effectively classical perturbations [30.20]. This particle production is due to the Gibbons–Hawking temperature in de Sitter space and is analogous to Hawking radiation from black holes.

These fluctuations determine the initial conditions for primordial density perturbations, in the subsequent radiation and dust-dominatederas. If inflation is driven by a slowly-rolling scalar field φ, then the

time-dependence of the field breaks the spacetime symmetry of de Sitter spacetime and defines a preferred time-like direction, and hence a foliation of spatial hypersurfaces orthogonal to $\nabla_\mu \varphi$. Quantum fluctuations perturb the classical evolution of the field and lead to adiabatic density perturbations as different parts of the universe reach the end of inflation, and reheat, having undergone different local expansion [30.21].

Any massless field can acquire a spectrum of quantum fluctuations on super-Hubble scales. Transverse and trace-free (tensor) metric perturbations correspond to gravitational waves, and inflation also stretches quantum fluctuations of the free gravitational field up to super-Hubble scales. These tensor perturbations could, in principle, be observed on large scales in the CMB sky today. Any detection of primordial gravitational waves on super-Hubble scales at last scattering would be strong evidence of a period of inflation in the very early universe. The power spectrum of these primordial gravitational waves would be a direct measure of the Gibbons–Hawking temperature during inflation.

Vector and scalar metric perturbations are related to vorticity and potential flows in matter fields. Vorticity usually decays rapidly in an inflating universe (and is identically zero for scalar fields) but scalar metric perturbations are naturally generated by quantum fluctuations in scalar fields, and massless fields acquire a power spectrum close to scale invariant in an almost de Sitter expansion. Small deviations from exact scale-invariance are expected [30.14] due to the slow change of the Hubble rate during inflation $\epsilon = -\dot{H}/H^2$, and the mass of the field relative to the Hubble scale $\eta = m^2/3H^2$. In the latest CMB data [30.16] there is now evidence of small, but significant deviations from exact scale-invariance at about the 1% level, consistent with the expectations of slow-roll inflation.

Further clues to the physical processes at work during a period of inflation in the very early universe could be revealed by other features in the distribution of primordial density perturbations and/or gravitational waves. Scale-dependence, nonadiabaticity, and non-Gaussianity of their distribution are all the subject of ongoing theoretical and observational work. The CMB currently provides the most detailed picture of primordial fields, but galaxy redshift surveys extend these constraints to smaller cosmic scales which will continue to improve with future surveys [30.22], while ambitious radio surveys with proposed experiments such as the SKA could map all the neutral hydrogen within our observable horizon via the intensity of the redshifted 21 cm line produced by the hyperfine splitting of the electron-proton ground state [30.23].

30.6 Outlook

Cosmology based on the FLRW metric (30.2) has emerged from hesitant beginnings, limited by sparse and uncertain data, to become one of the most active areas of modern science, with rich data sets and precise parameter constraints. Future observational projects will continue to drive the subject forward as astronomers accumulate and analyze large datasets, mapping out the visible matter within our horizon. Progress will need to go hand-in-hand with an improved understanding of the physical processes which lead to the formation of galaxies, stars, and black holes, with which we map out our universe.

Observational efforts have already led most cosmologists to conclude that the expansion of our universe is dominated by some form of dark energy, possibly a cosmological constant, causing the FLRW scale factor to accelerate at present. There is now intense effort to characterize this dark component [30.24] through its equation of state, or to test whether the observed acceleration could instead be due to modified gravity on large scales, or indeed nonlinear inhomogeneities which invalidate the FLRW description.

Observations of primordial density perturbations on scales larger than the horizon scale at the time of last-scattering of the CMB also suggest that our universe underwent a period of accelerated expansion, or inflation, at a much higher energy density in the very early universe. Quantum fluctuations about the FLRW background solution could then seed inhomogeneous structures in our primordial universe.

This description assumes a semiclassical approach where we consider quantum field fluctuations, including linear fluctuations in the metric, within a classical FLRW background. In the absence of a full theory of quantum gravity this seems a reasonable assumption, and is the same reasoning used to deduce Hawking radiation and the evaporation of black holes. However, the amplitude of these fluctuations is proportional to the Hubble scale, and becomes of the order of unity as the Hubble rate approaches the inverse Planck time,

$t_{\rm Pl}^{-1} = (5\times10^{-44}\,{\rm s})^{-1}$. At this point, the semiclassical description is expected to break down, and a more complete quantum gravity description is required, possibly resolving the singularities seen in classical FLRW solutions. Both Wheeler–De Witt quantization and loop quantum gravity [30.25] have also been developed largely within the context of FLRW spacetimes, due to their observational motivation and their analytical simplicity.

A different approach that has been advocated within the semiclassical picture is an eternally self-reproducing universe [30.26] where quantum fluctuations cause some spatial regions to jump to a higher density, despite the classical drift towards lower densities in an expanding universe. These higher density regions inflate faster than low density regions and the spatial volume on some constant time hypersurface could become dominated by regions that emerge from this self-reproducing regime. This suggests that on unobservably large scales the universe could be highly inhomogeneous, still undergoing high-energy inflation far beyond our horizon, and the FLRW metric (30.2) then becomes valid only on exceedingly large, but finite regimes, emerging at lower energies from this inhomogeneous quantum diffusion dominated regime.

FLRW spacetimes will continue to play a central role in both abstract theoretical speculations about our cosmological past and future, and in more pragmatic attempts to understand the distribution of matter and astrophysics in a cosmological neighborhood today.

References

30.1 S. Weinberg: *Cosmology* (Oxford Univ. Press, Oxford 2008)

30.2 N.J. Cornish, J.R. Weeks: Measuring the shape of the universe, Notices AMS **45**, 1463 (1998)

30.3 S. Cole, W.J. Percival, J.A. Peacock, P. Norberg, C.M. Baugh, C.S. Frenk, I. Baldry, J. Bland-Hawthorn, T. Bridges, R. Cannon, M. Colless, C. Collins, W. Couch, N.J.G. Cross, G. Dalton, V.R. Eke, R. De Propris, S.P. Driver, G. Efstathiou, R.S. Ellis, K. Glazebrook, C. Jackson, A. Jenkins, O. Lahav, I. Lewis, S. Lumsden, S. Maddox, D. Madgwick, B.A. Peterson, W. Sutherland, K. Taylor: The 2dF Galaxy Redshift Survey: Power-spectrum analysis of the final dataset and cosmological implications, Mon. Not. Roy. Astron. Soc. **362**, 505 (2005)

30.4 D.J. Eisenstein, et al.: SDSS-III: Massive spectroscopic surveys of the distant universe, the milky way galaxy, and extra-solar planetary systems, Astron. J. **142**, 72 (2011)

30.5 W.L. Freedman, B.F. Madore, B.K. Gibson, L. Ferrarese, D.D. Kelson, S. Sakai, J.R. Mould, R.C. Kennicutt, H.C. Ford, J.A. Graham, J.P. Huchra, S.M.G. Hughes, G.D. Illingworth, L.M. Macri, P.B. Stetson: Final results from the Hubble Space Telescope key project to measure the Hubble constant, Astrophys. J. **553**, 47 (2001)

30.6 S. Perlmutter, B.P. Schmidt: Measuring cosmology with supernovae, Lect. Notes Phys. **598**, 195 (2003)

30.7 M. Tegmark, et al.: Cosmological Constraints from the SDSS luminous red galaxies, Phys. Rev. D **74**, 123507 (2006)

30.8 H.S. Kragh: *Conceptions of Cosmos: From Myths to the Accelerating Universe. A History of Cosmology* (Oxford Univ. Press, Oxford 2007)

30.9 D.J. Fixsen, E.S. Cheng, J.M. Gales, J.C. Mather, R.A. Shafer, E.L. Wright: The Cosmic Microwave Background spectrum from the full COBE FIRAS data set, Astrophys. J. **473**, 576 (1996)

30.10 A.A. Penzias, R.W. Wilson: A Measurement of excess antenna temperature at 4080-Mc/s, Astrophys. J. **142**, 419 (1965)

30.11 G.F. Smoot, C.L. Bennett, A. Kogut, E.L. Wright, J. Aymon, N.W. Boggess, E.S. Cheng, G. De Amici, S. Gulkis, M.G. Hauser, G. Hinshaw, P.D. Jackson, M. Janssen, E. Kaita, T. Kelsall, P. Keegstra, C. Lineweaver, K. Loewenstein, P. Lubin, J.C. Mather, S.S. Meyer, S.H. Moseley, T. Murdock, L. Rokke, R.F. Silverberg, L. Tenorio, R. Weiss, D.T. Wilkinson: Structure in the COBE differential microwave radiometer first year maps, Astrophys. J. **396**, L1 (1992)

30.12 E. Komatsu, K.M. Smith, J. Dunkley, C.L. Bennett, B. Gold, G. Hinshaw, N. Jarosik, D. Larson, M.R. Nolta, L. Page, D.N. Spergel, M. Halpern, R.S. Hill, A. Kogut, M. Limon, S.S. Meyer, N. Odegard, G.S. Tucker, J.L. Weiland, E. Wollack, E.L. Wright: Seven-year Wilkinson microwave anisotropy probe (WMAP) observations: Cosmological interpretation, Astrophys. J. Suppl. **192**, 18 (2011)

30.13 G. Steigman: Primordial nucleosynthesis in the precision cosmology era, Annu. Rev. Nucl. Part. Sci. **57**, 463 (2007)

30.14 D.H. Lyth, A.R. Liddle: *The Primordial Density Perturbation: Cosmology, Inflation and the Origin of Structure* (Cambridge Univ. Press, Cambridge 2009)

30.15 J.M. Maldacena: The large N limit of superconformal field theories and supergravity, Adv. Theor. Math. Phys. **2**, 231 (1998)

30.16 G. Hinshaw, D. Larson, E. Komatsu, D.N. Spergel, C.L. Bennett, J. Dunkley, M.R. Nolta, M. Halpern, R.S. Hill, N. Odegard, L. Page, K.M. Smith, J.L. Weiland, B. Gold, N. Jarosik, A. Kogut, M. Limon,

S.S. Meyer, G.S. Tucker, E. Wollack, E.L. Wright: Nine-year Wilkinson microwave anisotropy probe (WMAP) observations: Cosmological parameter results, Astrophys. J. Suppl. **208**, 19 (2013)
30.17 K.A. Malik, D. Wands: Cosmological perturbations, Phys. Rep. **475**, 1 (2009)
30.18 J.M. Bardeen: Gauge invariant cosmological perturbations, Phys. Rev. D **22**, 1882 (1980)
30.19 T. Padmanabhan: *Structure Formation in the Universe* (Cambridge Univ. Press, Cambridge 1993)
30.20 C. Kiefer, D. Polarski: Why do cosmological perturbations look classical to us?, Adv. Sci. Lett. **2**, 164 (2009)
30.21 D. Wands, K.A. Malik, D.H. Lyth, A.R. Liddle: A New approach to the evolution of cosmological perturbations on large scales, Phys. Rev. D **62**, 043527 (2000)
30.22 L. Amendola, Euclid Theory Working Group: Cosmology and fundamental physics with the Euclid satellite, Living Rev. Rel. **16**, 6 (2013)
30.23 J.R. Pritchard, A. Loeb: 21-cm cosmology, Rept. Prog. Phys. **75**, 086901 (2012)
30.24 E.J. Copeland, M. Sami, S. Tsujikawa: Dynamics of dark energy, Int. J. Mod. Phys. D **15**, 1753 (2006)
30.25 M. Bojowald, C. Kiefer, P.V. Moniz: Proc. 12th Marcel Grossmann Meeting Gen. Relativ., Paris, France, July 12–18 (2009)
30.26 A.D. Linde: The selfreproducing inflationary universe, Sci. Am. **271**, 32 (1994)

31. Exact Approach to Inflationary Universe Models

Sergio del Campo

In this chapter we introduce a study of inflationary universe models that are characterized by a single scalar inflation field. The study of these models is based on two dynamical equations: one corresponding to the Klein–Gordon equation for the inflaton field and the other to a generalized Friedmann equation. After describing the kinematics and dynamics of the models under the Hamilton–Jacobi scheme, we determine in some detail scalar density perturbations and relic gravitational waves. We also introduce the study of inflation under the hierarchy of the slow-roll parameters together with the flow equations. We apply this approach to the modified Friedmann equation that we call the Friedmann–Chern–Simons equation, characterized by $\mathcal{F}(H) = H^2 - \alpha H^4$, and the brane-world inflationary models expressed by the modified Friedmann equation.

31.1 Aims and Motivations

The most appealing cosmological model to date is the standard hot big bang scenario. This model rests on the assumption of the cosmological principle that the universe is both homogeneous and isotropic at large scale [31.1–4]. Even though this model could explain observational facts such as approximately 3-K microwave background radiation [31.5], the primordial abundances of the light elements [31.6–8], the Hubble expansion [31.9, 10], and the present acceleration [31.11, 12], it presents some shortcomings (*puzzles*) when this is traced back to very early times in the evolution of the universe. Among them we distinguish the horizon, the flatness, and the monopole problems. In dealing with these *puzzles*, the standard big bang model demands an unacceptable amount of fine-tuning concerning the initial conditions for the universe.

Inflation has been proposed as a good approach for solving most of the cosmological *puzzles* [31.13, 14]. The essential feature of any inflationary universe model proposed so far is the rapid (accelerated) but finite period of expansion that the universe underwent at very early times in its evolution.

This brief accelerated expansion, apart of solving most of the cosmological problems mentioned above, serves to produce the seeds that, in the course of the subsequent eras of radiation and matter dominance, developed into the cosmic structures (galaxies and clusters thereof) that we observe today.

In fact, the present popularity of the inflationary scenario is entirely due to its ability to generate a spectrum of density perturbations which lead to structure formation in the universe. In essence, the conclusion

that all the observations of microwave background anisotropy performed so far support inflation rests on the consistency of the anisotropy with an almost Harrison–Zel'dovich power spectrum predicted by most of the inflationary universe scenarios [31.15].

The implementation of the inflationary universe model rests on the introduction of a scalar inflaton field ϕ. The evolution of this field becomes governed by its scalar potential $V(\phi)$ via the Klein–Gordon equation. Thus, this equation of motion together with the Friedmann equation, obtained from Einstein general relativity theory, form the most simple set of field equations, which could be applied to obtain inflationary solutions. However, in order to do this it is necessary to give an explicit expression for the scalar inflaton potential. However, in simple cases it is very complicated to find solutions, even in the situation that the so-called slow-roll approximation is applied, where the kinetics term is much smaller than the potential energy, i. e. $\dot{\phi}^2 \ll V(\phi)$, together with approximation $|\ddot{\phi}| \ll H|\dot{\phi}|$. From now on the dots represent the derivative with respect to the cosmological time t.

In general terms, the condition for inflation to occur is that the inflaton field slow rolls near the top of the potential for sufficiently long time, so that the vacuum energy drives the inflationary expansion of the universe. In this approach, many models of inflation have been proposed based on single-field or multifield potentials constructed in various theoretical schemes.

We may distinguish those solutions introduced by *Barrow* [31.16], where the scale factor $a(t)$ has the asymptotic property that ordinary differential equations of the form $\ddot{a} = P(a,t)/Q(a,t)$, as $t \longrightarrow \infty$ with polynomials P and Q, bring specific different solutions from which we can distinguish those named *logamediate* inflationary models. What is interesting in these models is that the property of the ratio of tensor to scalar perturbations is small and the power spectrum can be either red or blue tilted, according to the values of the parameters that characterize the model [31.17]. The motivation to study logamediate models comes from the form of the scalar potential that describes the model in the slow-roll approximation, i. e., $V(\phi) \propto \phi^{\alpha} \exp[-\beta\phi^{\gamma}]$, where α, β and γ are arbitrary constants. This potential includes exponential potential ($\alpha = 0$) that appears in Kaluza–Klein theories, as well as in supergravity, and in superstring models [31.18]. It also includes power-law potentials ($\beta = 0$) with models based on dynamical supersymmetry breaking which motivates potentials of the type $V(\phi) \propto \phi^{-\alpha}$ [31.19]. We should mention that one of the drawbacks of this model rests on the impossibility to bring inflation to an end. We know that at the end of the period of inflation the energy density of the universe is locked up in a combination of kinetic and potential energies of the scalar field, which drives inflation [31.20]. One path to defrost the universe after inflation is known as reheating [31.21]. During reheating most of the matter and radiation of the universe are created due to the decay of the inflaton field. While the temperature grows in many orders of magnitude, it is at this point where the universe matches the big bang model. In this process the particular interest is in the quantity known as the reheating temperature. This temperature is related to the temperature of the universe when the radiation epoch began.

In the reheating process, the oscillations of the inflaton are an essential part for the standard mechanism of reheating. However, in some models the inflaton potential does not present a minimum and, therefore, the scalar field does not oscillate. Here, the standard mechanism of reheating does not work [31.22]. These models are known in the literature as nonoscillating models [31.23, 24]. Nonoscillating models correspond to runaway fields such as module fields in string theory, which are potentially useful for inflation model building because they present flat directions which survive the famous η problem of inflation [31.25]. This problem is related to the fact that between the inflationary plateau and the quintessential tail there is a difference of over 100 orders of magnitude. There is a mechanism of reheating in this kind of model which is based on the introduction of the curvaton field [31.26]. The study of the curvaton reheating in a logamediate inflationary model was carried out in [31.27].

One way of finding inflationary solutions out of the slow-roll approximation is giving the functional form of the Hubble parameter in term of the inflaton field, i. e., $H(\phi)$, the so-called generating function [31.28, 29]. This approach presents some advantages when compared with the slow-roll approximation: first of all, the form of the potential is deduced, and second, since an exact solution is obtained, then, application to the final period of inflation is possible, where the kinetic term of the inflaton field in the Friedmann equation becomes important [31.30], i. e., during the reheating phase. The method followed here is usually referred to as the Hamilton–Jacobi (H–J) scheme [31.31–35].

There is a particular scenario of *intermediate inflation* [31.36, 37] in which the scale factor evolves

as $a(t) \approx \exp A\, t^f$, where A is constant and f is a free parameter which ranges as $0 < f < 1$. Therefore, the expansion of the Universe is slower than standard de Sitter inflation, but faster than power law inflation. The main motivation to study this latter kind of model came from string/M theory. Actually, the intermediate inflationary model was introduced as an exact solution for a particular scalar field potential of the type $V(\phi) \propto \phi^{-4(f^{-1}-1)}$. In the slow-roll approximation, and with this sort of potential, it is possible to have a spectrum of density perturbations which presents a scale-invariant spectral index, i. e., $n_s = 1$, the so-called Harrison–Zel'dovich spectrum provided that f takes the value of 2/3 [31.38–40]. Even though this kind of spectrum is disfavored by the current WMAP data, the inclusion of tensor perturbation, which could be presented at some point by inflation and parametrized by the tensor-to-scalar ratio r, the conclusion that $n_s \geq 1$ is allowed providing that the value of r is significantly nonzero. In fact, in [31.41] it was shown that the combination $n_s = 1$ and $r > 0$ is given by a version of the intermediate inflation in which the scale factor varies as $a(t) \propto e^{t^{2/3}}$ within the slow-roll approximation. We should mention here that, similar to the logamediate model, the resulting effective potential does not have a minimum, and therefore, the introduction of the curvaton field becomes necessary to bringing inflation to an end [31.42].

In this chapter we would like to study the consequences that result when considering a modified Friedmann equation expressed by

$$\mathcal{F}(H) \equiv \left(\frac{8\pi}{3m_{Pl}^2}\right)\rho_\phi = \left(\frac{8\pi}{3m_{Pl}^2}\right)\left[\frac{1}{2}\dot{\phi}^2 + V(\phi)\right]. \tag{31.1}$$

Here, $\mathcal{F} \geq 0$ is an arbitrary function of the Hubble parameter $H = \dot{a}/a$; a is the scale factor, and ρ_ϕ represents the scalar field energy density given by $\rho_\phi = 1/2\dot{\phi}^2 + V(\phi)$. Also, $V(\phi)$ expresses the scalar inflaton potential and $m_{Pl}^2 \equiv 1/G$ represents the Planck mass. The description of this chapter is based on a recent article by this author [31.43]. Here, we have added some topics for the sole purpose of completeness.

The motivation for using this kind of equation lies in the fact that in the literature the study of several models can be reduced to such a modified Friedmann equation. For instance, in an $L(R)$-theory of gravity in which $L(R) = R - \alpha^2/(3R)$, where R is the scalar curvature and α is a constant with dimension of mass square, the Friedmann equation becomes modified by the expression [31.44]

$$\mathcal{F}(H) = \frac{6H^2 - \frac{\alpha}{2}}{\frac{11}{8} - \frac{9}{4\alpha}H^2} .$$

Another example is the case in which the theories of generalized modified gravity present an acceleration equation of the second-order derivative. In this case, the Friedmann equation is written as [31.45]

$$H^2 + \frac{1}{6}f(H) - \frac{1}{6}Hf_{,H}(H) = \left(\frac{8\pi}{3m_{Pl}^2}\right)\rho_\phi , \tag{31.2}$$

where the function f is a function of the Hubble parameter H and $f_{,H} \equiv (\mathrm{d}f(H))/\mathrm{d}H$. In [31.45] different expressions were studied for the function f, giving different modified Friedmann equations.

It is also possible to consider

$$\mathcal{F}(H) = H^2 - \alpha H^4 = \left(\frac{8\pi}{3m_{Pl}^2}\right)\rho_\phi , \tag{31.3}$$

where α is a constant with a dimension of $mass^{-2}$.

There are various ways to obtain this latter expression for $\mathcal{F}(H)$. This has been derived by considering a quantum corrected entropy-area relation of the type [31.46]

$$S_{\mathcal{A}} = m_{Pl}^2\frac{\mathcal{A}}{4} - \tilde{\alpha}\ln\left(m_{Pl}^2\frac{\mathcal{A}}{4}\right), \tag{31.4}$$

where $\mathcal{A}$ is the area of the apparent horizon, and $\tilde{\alpha}$ is a dimensionless positive constant determined by the conformal anomaly of the fields. The conformal anomaly is interpreted as a quantum correction to the entropy of the apparent horizon [31.47]. In fact, in order to obtain (31.3), we consider the area expressed as $A = 4\pi r_A^2$, where r_A represents the apparent horizon, which for a flat universe becomes $r_A = 1/H$. Therefore, from (31.4) we obtain

$$\mathrm{d}S_{\mathcal{A}} = -2\pi m_{Pl}^2\frac{\mathrm{d}H}{H^3} + 2\tilde{\alpha}\frac{\mathrm{d}H}{H} . \tag{31.5}$$

From the first law of thermodynamics we have $T_A\,\mathrm{d}S_A = -\mathrm{d}E_A$, where T_A is the temperature of the ap-

parent horizon which is given by

$$T_A = \frac{1}{2\pi r_A} = \frac{H}{2\pi} ,$$

and

$$dE_A = \frac{4}{3}\pi r_A^3\, d\rho = \frac{4}{3}\pi \frac{d\rho}{H^3} ,$$

is the energy with ρ, the energy density. Thus, from these expressions, after an integration, we obtain (31.3), where the constant of integration has been chosen as zero, $\rho = \rho_\phi = \frac{1}{2}\dot{\phi}^2 + V(\phi)$, and $\alpha \equiv \tilde{\alpha}/(2\pi m_{Pl}^2)$.

On the other hand, the modified Friedmann equation, (31.3), could be obtained when an AdS-Schwarzschild black-hole via holographic renormalization is considered, together with mixed boundary conditions corresponding to the Einstein field equations in four dimensions [31.48]. Also, this could be derived in terms of spacetime thermodynamics together with a generalized uncertainty principle of quantum gravity [31.49]. A Chern–Simons type of theory also yields to this modification [31.50]. In this case, we will call the resulting modified Friedmann equation the Friedmann–Chern–Simons equation [31.43].

Superstring and M-theory bring about the possibility of considering our universe as a domain wall embedded in a higher dimensional space. In this scenario the standard model of particle physics is confined to the brane, while gravitation propagates into the bulk space–time. The effect of extra dimensions induces a change in the Friedmann equation. Here, the function $\mathcal{F}(H)$ results to be

$$\mathcal{F}(H) = \left(\frac{8\pi\lambda}{3m_{Pl}^2}\right)\left[\sqrt{1+\left(\frac{3m_{Pl}^2}{4\pi\lambda}\right)H^2} - 1\right] = \left(\frac{8\pi}{3m_{Pl}^2}\right)\rho_\phi , \tag{31.6}$$

where λ represents the brane tension [31.51–53].

Here, in this chapter, we first describe the inflationary paradigm within which we will depict a possible classification for different inflationary universe models. Then, after describing a general approach to the study of inflation based on the modified Friedmann equation within the scheme developed under the exact H–J approach, we determine in some detail scalar density perturbations and relic gravitational waves. We also introduce the study of inflation under the hierarchy of the slow-roll parameters together with the flow equations. In this context we will describe in some detail the latter two cases specified above, i. e., the Friedmann–Chern–Simons case, expressed by the modified Friedmann equation (31.3), and the brane-world inflationary universe models, described by (31.6).

31.2 Inflation as a Paradigm

As was mentioned above, the idea of inflation rests on a quasi exponential expansion that the universe underwent at an early time. In the period of 10^{-35} [s] a Planck-sized region blows up to a factor of 10^{30}!

However, the question is, what causes inflation? Essentially, it is assumed that inflation is produced, in its simplest version, by a scalar field ϕ usually called the inflaton field. This scalar field is characterized by its scalar field potential $V(\phi)$. In the slow-roll approximation for inflation the field equations become

$$H^2 \simeq \frac{8\pi}{3m_{Pl}^2} V(\phi) , \tag{31.7}$$

and

$$\dot{\phi}H \simeq -\frac{1}{3}\frac{dV(\phi)}{d\phi} = -\frac{1}{3}V'(\phi) , \tag{31.8}$$

where from now on one or more primes will represent derivatives with respect to the scalar field ϕ.

Thus, we see from this set of equations that for describing the physical evolution of the universe we need an explicit expression for the scalar field potential $V(\phi)$. In this context, many models for inflation depend explicitly on the form of the scalar potential. Actually, their functional forms characterize different inflationary models. For instance, the potential shown in Fig. 31.1 corresponds to *old inflation* [31.13].

This kind of model is characterized by a first-order phase transition, where percolation becomes an essential issue. Unfortunately, this model cannot realize the appropriate amount of inflation needed for solving most of the puzzles present in standard big bang theory. In order to solve the problem presented in the old inflationary model, another sort of model emerged, so-called

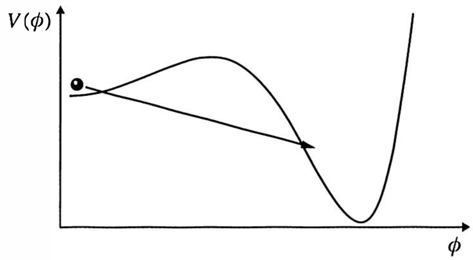


Fig. 31.1 Typical form of the scalar potentials $V(\Phi)$ as a function of the inflaton field ϕ in the case of *old inflation*

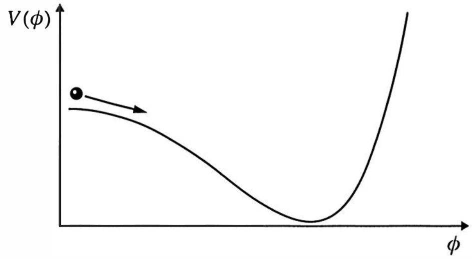


Fig. 31.2 Typical form of the scalar potentials $V(\Phi)$ as a function of the inflaton field ϕ in the case of *new inflation*

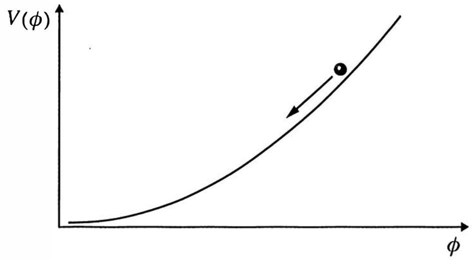


Fig. 31.3 Typical form of the scalar potentials $V(\Phi)$ as a function of the inflaton field ϕ in the case of *chaotic inflation*

new inflation [31.55, 56]. Differently from in the old inflationary model, the Universe underwent a second-order phase transition at an early time. The scalar field potential form is shown in Fig. 31.2.

Then another quite different sort of model appeared, so-called *chaotic inflation* [31.57]. The main difference from the previous models is that in this case there is no need for any kind of phase transition. In fact, inflation may start with a value for the inflaton field so high that it may exceed the value of the Planck mass. Figure 31.3

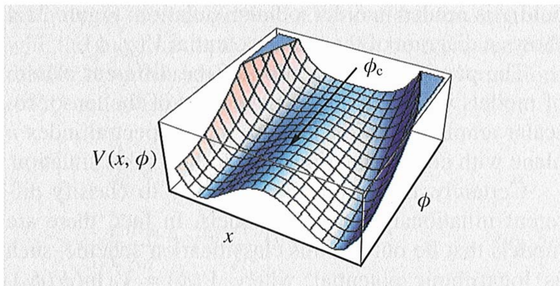


Fig. 31.4 Form of the scalar potential $V(\chi, \phi)$ as a function of the scalar fields,χ and ϕ, for the case of *hybrid inflation*, where the scalar potential has the $V(\chi, \phi) = V_0 + \frac{1}{2}g^2(\phi^2 - \phi_c^2)\chi^2 + \frac{1}{2}m^2\phi^2 + \frac{1}{4}\lambda\chi^4$ form

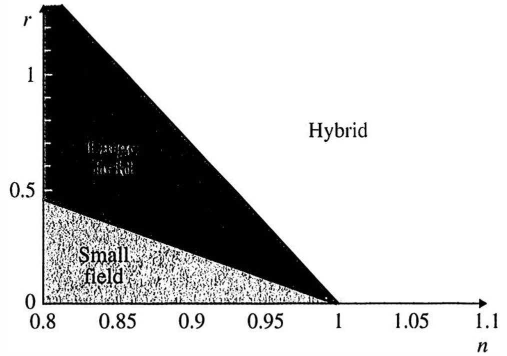


Fig. 31.5 Possible classification for the different inflationary universe models (after [31.54])

shows the form of the potential for a chaotic inflationary universe model.

In general terms, it is possible to classify the different inflationary models into three categories [31.58]. The first category is the small field (like old or new inflation). Here, the inflaton potential $V(\phi)$ has the constraints $V'' < 0$ and $(\lg V)'' < 0$. The second category is the *large field* (like chaotic inflation). Now, the inflaton potential $V(\phi)$ has the constraints $V'' > 0$ and $(\lg V)'' < 0$. The third category corresponds to the *hybrid models* [31.59]. In this latter case two scalar fields act, and the form of the scalar potential looks like

$$V(\chi, \phi) = V_0 + \frac{1}{2}g^2\left(\phi^2 - \phi_c^2\right)\chi^2 + \frac{1}{2}m^2\phi^2 + \frac{1}{4}\lambda\chi^4 ,$$

where V_0, g, ϕ_c and λ are constant. Note that here the effective squared mass of the scalar field χ is $g^2(\phi^2 - \phi_c^2)$, where ϕ corresponds to the inflaton field, and the scalar

field χ is needed in order to finish inflation. Figure 31.4 shows a diagram of the scalar potential $V(\chi, \phi)$.

The previous classification of the different classes of models will cover different regions of the tensor-to-scalar amplitude ratio r and the scalar spectral index n plane with no overlap. Figure 31.5 shows this situation.

Certainly, this is not the only way to classify different inflationary universe models. In fact, there are models that lie outside this classification scheme, such as logarithmic potential, where $V(\phi) = V_0 \ln(\phi/\phi_0)$. Also, we could add the *medium* field classification, which could be large, medium, or small. These refer to field variations which are much larger than, comparable to, or much smaller than the Planck mass.

Note that, in general terms, large (or medium) field inflation is difficult to describe from the fundamental point of view without a complete theory of quantum gravity. On the other hand, small field inflation does not seem to have this problem, but still has to deal with the problem related to the onset of inflation, i. e., what was the state of the universe prior to the period of inflation? Answers are highly dependent on the initial conditions that the universe presents at that epoch, which complicates any model predictions.

31.3 The Exact Solution Approach

Let us consider (31.1) together with the evolution equation related to the scalar inflaton field ϕ in a background of a flat Friedmann universe

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = 0 . \tag{31.9}$$

In obtaining this latter equation we have assumed that the matter, specified by the inflaton scalar field, enters into the Lagrangian action in such a way that its variation in a Friedmann–Robertson–Walker background metric leads to the Klein–Gordon equation expressed by (31.9). Therefore, we are considering constrained sorts of models, in which the background (together with the perturbed equation, see (31.37)) is not modified. The theory of gravity , such as Hořava-Lifshitz [31.60], lies outside of the approach followed here. Also, in this study we will use the scalar field ϕ as a *time variable*. The requirement imposed in this approach is that the scalar field increases monotonically and its time derivative $\dot{\phi}$ should not change sign along the path of evolution.

It is not hard to find that

$$\dot{\phi} = -\left(\frac{m_{\text{Pl}}^2}{8\pi}\right) \mathcal{F}_{,H} \left(\frac{H'}{H}\right), \tag{31.10}$$

where, as before, $\mathcal{F}_{,H} \equiv \mathrm{d}\mathcal{F}/\mathrm{d}H$. From this expression we obtain the scalar potential results

$$V(\phi) = \frac{3m_{\text{Pl}}^2}{8\pi}\mathcal{F}\left[1 - \frac{m_{\text{Pl}}^2}{48\pi}\left(\frac{H'\mathcal{F}_{,H}}{H\sqrt{\mathcal{F}}}\right)^2\right]. \tag{31.11}$$

It is not hard to show that

$$aH = -\left(\frac{m_{\text{Pl}}^2}{8\pi}\right)\frac{\mathcal{F}_{,H}}{H}a'H' , \tag{31.12}$$

from which we obtain

$$a(\phi) = a_{\text{i}} \exp\left(-\frac{8\pi}{m_{\text{Pl}}^2}\int_{\phi_{\text{i}}}^{\phi}\frac{H^2}{H'\mathcal{F}_{,H}}\mathrm{d}\phi\right). \tag{31.13}$$

where $a_{\text{i}} = a(\phi_{\text{i}})$.

It is not hard to see that the acceleration equation for the scale factor is

$$\frac{\ddot{a}}{a} = H^2(1 - \epsilon_{\text{H}}) , \tag{31.14}$$

where the function ϵ_{H} corresponds to

$$\epsilon_{\text{H}} \equiv -\frac{\mathrm{d}\ln H}{\mathrm{d}\ln a} = \left(\frac{m_{\text{Pl}}^2}{8\pi}\right)\frac{\mathcal{F}_{,H}}{H}\left(\frac{H'}{H}\right)^2 . \tag{31.15}$$

From this latter expression we can see that this definition, called the *first Hubble slow-roll parameter*, gives information about the acceleration of the universe. During inflation we have $\epsilon_{\text{H}} < 1$, and this period ends when ϵ_{H} takes the value equal to 1. In the next section we will use this parameter for describing scalar and tensor perturbations.

One interesting quantity in characterizing inflationary universe models is the amount of inflation. Usually, this quantity is defined by

$$N(t) \equiv \ln\frac{a(t_{\text{end}})}{a(t)} , \tag{31.16}$$

where $a(t_{\text{end}})$ corresponds to the scale factor evaluated at the end of inflation. In terms of the scalar field, to-

gether with the modified Friedmann equation we obtain

$$N(\phi) = \int_{t}^{t_{\rm end}} H \mathrm{d}t = \left(\frac{8\pi}{m_{\rm Pl}^2}\right) \int_{\phi_{\rm end}}^{\phi} \frac{H^2}{H' \mathcal{F}_{,H}} \mathrm{d}\phi$$

$$= \int_{\phi_{\rm end}}^{\phi} \frac{1}{\epsilon_{\rm H}} \frac{H'}{H} \mathrm{d}\phi \,. \tag{31.17}$$

Here, $\phi_{\rm end}$ represents the value of the scalar field at the end of inflation. Its value is determined by imposing $\epsilon_{\rm H}(\phi_{\rm end}) = 1$.

However, it seems to be more appropriate to describe the amount of inflation in terms of the comoving Hubble length, $1/(aH)$ than in terms of the scale factor only. In this case, the amount of inflation becomes [31.61]

$$\overline{N} = \ln \frac{a\,(t_{\rm end})\, H\,(t_{\rm end})}{a(t) H(t)} \,, \tag{31.18}$$

which results into

$$\overline{N}(\phi) = \int_{\phi_{\rm end}}^{\phi} \left(\frac{1}{\epsilon_{\rm H}} - 1\right) \frac{H'}{H} \mathrm{d}\phi \,. \tag{31.19}$$

Note that, in general, $\overline{N}(\phi)$ is smaller than $N(\phi)$ and they coincide only in the slow-roll limit. As was stressed in [31.61], we should consider $\overline{N}(\phi)$ and not $N(\phi)$ in determining the appropriate amount of inflation.

It is not enough to see that a given inflationary universe model presents an accelerated phase with a given number of e-folds associated to this period, as we did previously, but it is also necessary to show that their solutions are independent from their initial conditions. This ensures the true predictive power that presents any inflationary universe model, otherwise the corresponding physical quantities associated with the inflationary phase, such as the scalar or tensor spectra, would depend on these initial conditions. Thus, with the purpose of being predictive, every inflationary model needs its solutions to present an attractor behavior, in the sense that solutions with different initial conditions tend to a unique solution [31.62].

In order to study the corresponding inflationary attractor solution for our case, and following [31.62], we consider a linear perturbation $\delta H(\phi)$ around a given solution, expressed by $H_0(\phi)$. Below we will refer to this quantity as the background solution, and any quantity with the subscript zero is assumed to be evaluated with this background solution. Thus, from the field equations (31.1) and (31.9) we have

$$\left[1 + \frac{1}{3}\epsilon_{\rm H}\left(1 - H\frac{\mathcal{F}_{,HH}}{\mathcal{F}_{,H}}\right)\right]\bigg|_0 \delta H$$

$$\simeq \frac{1}{3}\left(\frac{m_{\rm Pl}^2}{8\pi}\right) \mathcal{F}_{,H} \frac{H'}{H^2}\bigg|_0 \delta H' \,, \tag{31.20}$$

This latter expression can be solved by obtaining

$$\delta H(\phi) = \delta H(\phi_{\rm i}) \exp \int_{\phi_{\rm i}}^{\phi} \left(\frac{3}{\epsilon_{\rm H}}\right)$$

$$\times \left[1 + \frac{1}{3}\epsilon_{\rm H}\left(1 - H\frac{\mathcal{F}_{,HH}}{\mathcal{F}_{,H}}\right)\right] \frac{H'}{H}\bigg|_0 \mathrm{d}\phi \,, \tag{31.21}$$

where $\phi_{\rm i}$ corresponds to some arbitrary initial value of ϕ. By considering theories in which $\mathcal{F}_{,H} > H\mathcal{F}_{,HH}$ we find that the integrand within the exponential term will be negative, since $\mathrm{d}\phi$ and H' have opposite signs (assuming that $\dot{\phi}$ does not change sign due to the perturbation δH)[31.61]. Therefore, all linear perturbations tend to vanish quickly.

On the other hand, we can show that the set of initial conditions is quite large by considering the dynamical system approach for inflationary universe models. In this respect we consider the field equations, from which it is not hard to show that

$$\frac{\mathrm{d}\dot{\phi}}{\mathrm{d}\phi} = -\left(\frac{m_{\rm Pl}^2}{8\pi}\right)$$

$$\times \left[[H\mathcal{F}_{,HH} - \mathcal{F}_{,H}]\left(\frac{H'}{H}\right)^2 + \mathcal{F}_{,H}\left(\frac{H''}{H}\right)\right]. \tag{31.22}$$

From this expression we obtained the phase diagram in the $\dot{\phi}-\phi$ plane. For the standard case in which $\mathcal{F} = H^2$, the above equation (31.22) reduces to

$$\frac{\mathrm{d}\dot{\phi}}{\mathrm{d}\phi} = -\left(\frac{m_{\rm Pl}^2}{8\pi}\right) H''.$$

or equivalently,

$$\frac{\mathrm{d}\dot{\phi}}{\mathrm{d}\phi} = -\frac{3\dot{\phi}\sqrt{\left(\frac{8\pi}{3m_{\rm Pl}^2}\right)\left(\frac{1}{2}\dot{\phi}^2 + V(\phi)\right)} + V'(\phi)}{2\dot{\phi}} \,,$$

which, for the case where the scalar potential corresponds to a massive scalar field $V(\phi) = \frac{1}{2}m^2\phi^2$ the latter equation reduces to

$$\frac{d\dot{\phi}}{d\phi} = -\frac{\sqrt{\frac{12\pi}{m_{Pl}^2}}\dot{\phi}\sqrt{\dot{\phi}^2 + m^2\phi^2} + m^2\phi}{2\dot{\phi}}.$$

Figure 31.6 shows the phase portrait for the case discussed above. This diagram shows the important feature related to the existence of an attractor solution to which all other solutions converge in time.

Before concluding this section, we should mention that there is another way of studying the scalar inflaton field equation (31.9), which could be written in terms of the *number of e-folds N* or, in terms of the *modified number of e-folds*, $\overline{N}$. Let us consider the latter one; then, (31.9) becomes

$$\left(\frac{(1-\epsilon_H)^2}{(1-\epsilon_H)(\epsilon_H - 3) + \frac{1}{H}\dot{\epsilon}_H}\right)\frac{d^2\phi}{d\overline{N}^2} + \frac{d\phi}{d\overline{N}} + \frac{1}{H^2}\left(\frac{1}{(1-\epsilon_H)(\epsilon_H - 3) + \frac{1}{H}\dot{\epsilon}_H}\right)\frac{dV}{d\phi} = 0, \tag{31.23}$$

where we used that $d/dt \equiv -H(1-\epsilon_H)\,d/d\overline{N}$. Also, in this equation the scalar field potential becomes

$$V(\phi) = \left(\frac{3m_{Pl}^2}{8\pi}\right)\mathcal{F}(H)\left(1 - \frac{1}{6}\frac{d\ln\mathcal{F}(H)}{d\ln H}\epsilon_H\right). \tag{31.24}$$

Since $\mathcal{F}(H) > 0$ and because ρ_ϕ should be positive, we then need to satisfy the inequality

$$\epsilon_H > 6\left(\frac{d\ln\mathcal{F}(H)}{d\ln H}\right),$$

in order for the scalar field potential to become positive.

In the slow-roll approximation, where $\epsilon_H = \epsilon \ll 1$, together with $H^2 \approx V(\phi)$ for the standard case, (31.23) simplifies to

$$\frac{d^2\phi}{d\overline{N}^2} + \left(\frac{\dot{\epsilon}}{H} - 3\right)\frac{d\phi}{d\overline{N}} + \frac{d}{d\phi}(\ln V) = 0.$$

In the case in which the scalar potential is taken to be a massive scalar field, i. e., $V(\phi) = \exp(\frac{1}{2}m^2\phi^2)$, the above equation represents an oscillator damped by the factor $\dot{\epsilon}/H - 3 > 0$, making its evolution slower. An interesting result is the situation in which $\dot{\epsilon}/H < 3$. Here, the oscillations are undamped; a situation that deserves to be studied more deeply.

Below we will describe some examples where the generating function $H(\phi)$ has been given explicitly. For instance, in the intermediate inflationary universe model already described in Sect. 31.1 [31.64], $H(\phi)$ is taken to be

$$H(\phi) = \frac{1}{2}f\beta^{\beta/4}\left(\frac{m_{Pl}^2}{2\pi}A\right)^{\beta/4}\phi^{-\beta/2}, \tag{31.25}$$

where $A > 0$ is a constant, f is in the range $0 < f < 1$, and $\beta = 4((1/f) - 1)$. The corresponding scalar potential becomes

$$V(\phi) = \frac{m_{Pl}^4}{(4\pi)^2}\frac{8A^2}{(\beta+4)^2}\left(\frac{2\pi\phi^2}{m_{Pl}^2\beta}\right)^{-\beta/2}\left(6 - \frac{\beta^2}{\phi^2}\right), \tag{31.26}$$

and the scale factor results in

$$a(t) = a_0 \exp At^f. \tag{31.27}$$

The scalar potential, (31.26) should be compared with the case where the slow-roll approximation is used. In this approximated case we obtain

$$V(\phi) \simeq \frac{m_{Pl}^4}{(4\pi)^2}\frac{48A^2}{(\beta+4)^2}\left(\frac{2\pi\phi^2}{m_{Pl}^2\beta}\right)^{-\beta/2}, \tag{31.28}$$

which is appropriated for the case $\phi \gg \beta/\sqrt{6}$ in expression (31.26).

Another example corresponds to the case where Einstein's equation is considered in such a way that inflation is driven by the evolution of scalar fields in potentials which possess minima [31.65]. Here, the generating function is taken to be

$$H(\phi) = \lambda A^2 \cosh\left(\frac{\phi}{A}\right), \tag{31.29}$$

where λ and A are two constants. The scalar potential becomes

$$V(\phi) = (\lambda A)^2\left[(3A^2 - 2)\cosh^2\left(\frac{\phi}{A}\right) + 2\right]. \tag{31.30}$$

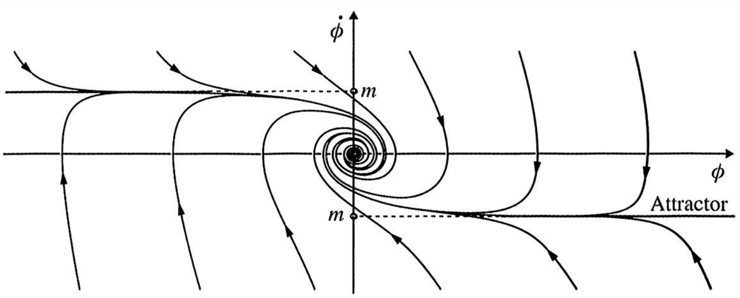


Fig. 31.6 The phase portrait for the case in which the scalar potential corresponds to $V(\phi) = (1/2)m^2\phi^2$, showing the attractor inflationary solution (after [31.63])

As a function of the cosmological time the scalar field and the scale factor become

$$\phi(t) = A \ln[\tanh(\lambda t)] , \tag{31.31}$$

and

$$a(t) = a_0 [\sinh(2\lambda t)]^{\frac{A^2}{2}} , \tag{31.32}$$

respectively. Note that we could have three cases for the constant A, namely, $3A^2 > 2$, $3A^2 < 2$, and $3A^2 = 2$. From these three cases, the first one is the most interesting, where the scalar potential has the property of being concave with a single stable minimum located at $\phi = 0$. The behavior of the universe at small t becomes $a(t) \approx t^{A^2/2}$, which corresponds to a power-law inflation for $A^2 > 2$, and the potential becomes

$$V(\phi) \approx \exp\left(\frac{2\phi}{A}\right)$$

for $\phi \longrightarrow -\infty$.

Another situation in which an explicit expression for the Hubble parameter has been used is in [31.62]. In this reference,

$$H(\phi) = H_0 \exp(-\beta\phi) , \tag{31.33}$$

where the constant H_0 becomes

$$H_0 = \left(\frac{8\pi V_0}{3m_{\mathrm{Pl}}^2}\frac{1}{1-1/3p}\right)^{1/2}$$

and

$$\beta = \frac{1}{m_{\mathrm{Pl}}}\left(\frac{4\pi}{p}\right)^{1/2} ,$$

with $V_0 > 0$ and $p > 1$ in order for inflation to occur. In this case the scalar potential becomes exponential, i. e.,

$$V(\phi) = V_0 \exp(-2\beta\phi) .$$

The scalar field as a function of time becomes

$$\phi(t) = m_{\mathrm{Pl}}\sqrt{\frac{p}{4\pi}} \ln\left[\sqrt{\frac{8\pi V_0}{3m_{\mathrm{Pl}}^2 p^2}\frac{1}{1-1/3p}}\,t\right] ,$$

and the scale factor corresponds to a power law and is given simply by $a(t) = a_0 t^p$.

In a gravity rainbow theory [31.66–68], where $\mathcal{F}(H) = f^2H^2$, with f a correction term that correlates with the energy of the probe particles [31.69],

$$H(\phi) = H_0\frac{\phi^M}{f^2}$$

was used, where H_0 and M are numbers greater than 1. It was found that the scalar field and the scale factor become

$$\phi(t) = \begin{cases} \left[\phi_0^{2-M} + \left(\frac{H_0 M(M-2)m_{\mathrm{Pl}}^2}{4\pi}\right)t\right]^{1/(2-M)}, & M \neq 2 \\ \phi_0 \exp\left[-\left(\frac{H_0 m_{\mathrm{Pl}}^2}{2\pi}\right)t\right], & M = 2 \end{cases}$$

and

$$a(\phi) = \begin{cases} a_0 \exp\left[\left(\frac{2\pi}{Mf^2 m_{\mathrm{Pl}}^2}\right)(\phi_0^2 - \phi^2)\right], & M \neq 2 \\ a_0 \exp\left[\left(\frac{\pi}{f^2 m_{\mathrm{Pl}}^2}\right)(\phi_0^2 - \phi^2)\right], & M = 2 \end{cases}$$

respectively. In this case, the scalar potential was

$$V(\phi) = \left(\frac{3H_0^2 m_{\mathrm{Pl}2}}{8\pi}\right)\phi^{2M} \times \left[1 - \left(\frac{M^2 f^2 m_{\mathrm{Pl}}^2}{12\pi}\right)\left(\frac{1}{\phi^2}\right)\right] .$$

An exact solution to Einstein's equations that describe the evolution of the cosmological chaotic inflationary universe model was presented in [31.70].

The generating function used in this case was $H(\phi) = \alpha_1\phi^2 + \alpha_2$, where α_1 and α_2 are two positive constants. The scalar field and the scale factor are

$$\phi = \phi_0 \exp\left[-\left(\frac{\alpha_1 m_{\mathrm{Pl}}^2}{2\pi}\right) t\right]$$

and

$$a(t) = a_0 \exp\left\{\alpha_2 t + \left(\frac{\pi\phi_0^2}{m_{\mathrm{Pl}}^2}\right) \times \left(1 - \exp\left[\left(\frac{\alpha_1 m_{\mathrm{Pl}}^2}{\pi}\right) t\right]\right)\right\}.$$

The scalar potential assumes the following form

$$V(\phi) = \frac{\lambda}{8}\left(\phi^2 - \nu^2\right)^2,$$

where $\lambda = (3\alpha_1^2 m_{\mathrm{Pl}}^2)/\pi$ and $\nu^2 \equiv \alpha_1/\alpha_2$.

Finally, let us consider the *logamediate* inflationary universe model [31.16] where the scale factor becomes

$$a(t) = \exp\left[A(\ln t)^\lambda\right],$$

with $t > 1$. Here $A > 0$ and $\lambda > 1$ are two constants. The Hubble parameter as a function of t is

$$H(t) = \frac{A\lambda(\ln t)^{\lambda-1}}{t},$$

where $A\lambda > 0$ in order to have an expanding universe. We should notice that when $\lambda = 1$, the model reduces to the power-law inflation, where $a(t) \approx t^p$, with $p = A > 1$.

The scalar field ϕ as a function of the cosmological time becomes

$$\phi(t) = \phi_0 + \gamma\sqrt{A\lambda}(\ln t)^{1/\gamma},$$

where γ corresponds to $\gamma = \frac{2}{\lambda+1}$.

The scalar field potential becomes (by setting $\phi_0 = 0$)

$$V(\phi) = V_0\phi^\alpha \exp\left(-\beta\phi^\gamma\right) \times \left\{1 - \frac{m_{\mathrm{Pl}}^2}{24\pi}\left(\frac{\gamma}{\phi}\right)^2\left[1 - \left(\frac{1}{\sqrt{A\lambda}}\right)^\gamma \phi^\gamma\right]\right\}, \tag{31.34}$$

where $\alpha = 4(\lambda - 1)/(\lambda + 1)$ and $\beta = 2((\lambda + 1)/(2\sqrt{A\lambda}))^\gamma$. Under the slow-roll approximation the scalar potential (31.34) becomes [31.17]

$$V(\phi) \simeq V_0\phi^\alpha \exp\left(-\beta\phi^\gamma\right),$$

which occurs in the range

$$\frac{m_{\mathrm{Pl}}}{\sqrt{6\pi}}\left(\frac{1}{\lambda+1}\right) \ll \phi \ll \sqrt{A\lambda}$$

of the scalar inflaton field.

31.4 Scalar and Tensor Perturbations

Inflation generates perturbations through the amplification of quantum fluctuations, which are stretched to astrophysical scales by brief, but rapid inflationary expansion. The simplest models of inflation generate two types of perturbations, density perturbations which come from quantum fluctuations of the scalar field [31.71–74], and relic gravitational waves which are tensor metric fluctuations [31.75–79]. The former experience gravitational instability leading to structure formation [31.80, 81], while the latter predict a stochastic background of relic gravitational waves which could influence the cosmic microwave background anisotropy via the presence of polarization [31.82, 83]. Upcoming experiments such as the Planck satellite will characterize polarization anisotropy to a higher accuracy [31.84]. It is very timely to develop the tools which can optimally utilize the polarization information to constrain models of the early Universe. Specifically, magnetic modes (B-modes) are signals from cosmic inflation and suggest the presence of gravitational waves [31.85].

In order to describe these perturbations, let us consider the *Hamilton–Jacobi* slow-roll parameters. The *first Hubble slow-roll parameter* ϵ_{H} was already defined in the previous section, (31.15).

The *second slow-roll parameter* η_{H} is defined as

$$\eta_{\mathrm{H}} \equiv -\frac{\mathrm{d}\ln H'}{\mathrm{d}\ln a} = \left(\frac{m_{\mathrm{Pl}}^2}{8\pi}\right)\frac{\mathcal{F}_{,H}}{H}\frac{H''}{H}. \tag{31.35}$$

Part E | 31.4

We should note here that both ϵ_H and η_H are exact quantities, despite the fact that we call them *slow-roll parameters*. In the slow-roll limit these parameters become [31.61, 86]

$$\epsilon_H \longrightarrow \epsilon\,, \qquad \eta_H \longrightarrow \eta - \epsilon\,, \tag{31.36}$$

where the quantities ϵ and η are common *slow-roll parameters* which satisfy $\epsilon \ll 1$ and $\eta \ll 1$, in agreement with the slow-roll approximation.

The evolution equation for the Fourier modes of the scalar perturbations (quantum mode functions) at some comovil wave number scale k is governed by [31.87–90]

$$\frac{d^2 u_k}{d\tau^2} + \left(k^2 - \frac{1}{z}\frac{d^2 z}{d\tau^2}\right) u_k = 0\,, \tag{31.37}$$

where τ represents the conformal time defined by $\tau = \int (1/a)\,dt$ and u_k corresponds to the Fourier transform of the Mukanov variable, which is defined by $u = z\mathcal{R}$, with $z = a\dot{\phi}/H$ and $\mathcal{R}$ defining the gauge-invariant comovil curvature perturbation. This latter amount remains constant outside the horizon, i. e., metric perturbations with wavelengths larger than the Hubble radius will be frozen [31.91].

During inflation it is expected that

$$k^2 \gg \frac{1}{z}\frac{d^2 z}{d\tau^2}\,,$$

i. e., the physical modes are assumed to have a wavelength much smaller than the curvature scale, and thus (31.37) can be solved to achieve

$$u_k(\tau) \approx e^{-ik\tau}\left(1 + \frac{\mathcal{A}_k}{\tau} + \ldots\right). \tag{31.38}$$

On the other hand, when

$$k^2 \ll \frac{1}{z}\frac{d^2 z}{d\tau^2}\,,$$

we have that the physical modes correspond to wavelengths much bigger than the curvature scale.

The mass term

$$\frac{1}{z}\frac{d^2 z}{d\tau^2}\,,$$

in our case becomes

$$\begin{aligned}\frac{1}{z}\frac{d^2 z}{d\tau^2} = 2a^2H^2\Bigg[&1 - \frac{3}{2}\eta_H + \frac{1}{2}\epsilon_H\left(5 - 3H\frac{\mathcal{F}_{,HH}}{\mathcal{F}_{,H}}\right)\\ &+\epsilon_H^2\left(1 - H\frac{\mathcal{F}_{,HH}}{\mathcal{F}_{,H}} + \frac{1}{2}H^2\frac{\mathcal{F}_{,HHH}}{\mathcal{F}_{,H}}\right)\\ &+\frac{1}{2}\eta_H^2 - \frac{1}{2}\epsilon_H\eta_H\left(3 - 2H\frac{\mathcal{F}_{,HH}}{\mathcal{F}_{,H}}\right)\\ &+\frac{1}{2H}\dot{\epsilon}_H\left(3 - 2H\frac{\mathcal{F}_{,HH}}{\mathcal{F}_{,H}}\right) - \frac{1}{2H}\dot{\eta}_H\Bigg]\,.\end{aligned} \tag{31.39}$$

For $\mathcal{F}(H) = H^2$ we obtain [31.92]

$$\begin{aligned}\frac{1}{z}\frac{d^2 z}{d\tau^2} = 2a^2H^2\Bigg(&1 + \epsilon_H - \frac{3}{2}\eta_H - \frac{1}{2}\epsilon_H\eta_H\\ &+\frac{1}{2}\eta_H^2 + \frac{1}{2H}\dot{\epsilon}_H - \frac{1}{2H}\dot{\eta}_H\Bigg)\,.\end{aligned}$$

It has long been known that (31.37) can be solved exactly in the case in which the mass term $(1/z)(d^2z/d\tau^2)$ is proportional to τ^{-2}, where this equation reduces to a Bessel equation, where the standard solution becomes $u_k \approx \sqrt{-k\tau}H_\nu(-k\tau)$, with H_ν the Hankel function of first kind, and the parameter ν depending on the slow-roll parameter ϵ via $\nu = 3/2 + \epsilon/(1-\epsilon)$. For instance, this occurred in the case of the standard Friedmann equation and the scale factor $a(t)$ expands as a power law, i. e., $a(t) \approx t^p\,(p > 1)$, resulting in $\epsilon_H = \eta_H =$ constant [31.93]. Other solutions, which are far from the slow-roll approximation, are described in [31.94].

Immediately that we obtain an explicit expression for u_k we can obtain the power spectrum, which is defined in terms of the two-point correlation function as

$$\mathcal{P}_\mathcal{R}(k) = \frac{k^3}{2\pi^2}\left\langle \mathcal{R}_{\vec{k}}, \mathcal{R}_{\vec{k}}\right\rangle \delta\left(\vec{k}' + \vec{k}\right), \tag{31.40}$$

which in terms of u_k and z becomes

$$\mathcal{P}_\mathcal{R}(k) = \frac{k^3}{2\pi^2}\left|\frac{u_k}{z}\right|^2\,. \tag{31.41}$$

In order to obtain u_k by solving (31.37), we need to impose some boundary conditions. These asymptotic conditions are usually taken to be the so-called Bunch–Davies vacuum states [31.95]

$$u_k \to \begin{cases}\frac{1}{\sqrt{2k}}e^{-ik\eta} & \text{as } -k\eta \longrightarrow \infty\,,\\ \mathcal{A}_k z & \text{as } -k\eta \longrightarrow 0\,.\end{cases} \tag{31.42}$$

This ensures that perturbations that are generated well inside the horizon, i. e., in the region where $k \ll aH$, the modes approach plane waves and those that are generated well outside the horizon, i. e., in the region where $k \gg aH$, remain unchanged.

The description of the primordial curvature perturbation presents a standard result given by [31.73, 74, 91, 96]

$$\mathcal{P}_{\mathcal{R}}(k) = \left(\frac{H}{|\dot{\phi}|}\right)^2 \left(\frac{H}{2\pi}\right)^2 \Bigg|_{aH=k} . \quad (31.43)$$

This perturbation is, in general, a function of the wave number k, which is evaluated as $aH = k$, i. e., when a given mode crosses outside the horizon during inflation. Since the modes do not evolve outside the horizon, the amplitude of the modes when they cross back inside the horizon coincides with the value that they had when they left the horizon.

By using the primordial scalar perturbations we can introduce the scalar spectral index n_s defined by

$$n_s - 1 \equiv \frac{d \ln \mathcal{P}_{\mathcal{R}}}{d \ln k} . \quad (31.44)$$

This quantity becomes

$$n_s - 1 = 2\eta_H - 2\left(3 - H\frac{\mathcal{F}_{,HH}}{\mathcal{F}_{,H}}\right)\epsilon_H . \quad (31.45)$$

Note that $n_s > 1$ requires

$$\eta_H > \left(3 - H\frac{\mathcal{F}_{,HH}}{\mathcal{F}_{,H}}\right)\epsilon_H ,$$

which corresponds to a *blue* spectral [31.97]. In the special case where $\mathcal{F}(H) = H^2$, we find $\eta_H > 2\epsilon_H$, and since ϵ_H by definition is positive, then, at the lowest order, $\eta_H > 2\epsilon_H > 0$. As was mentioned in [31.97], this condition is not easy to satisfy and this is particularly so during the final stage of inflation where $\epsilon \simeq 1$, which requires that $\eta > 2$.

In the same way we define the *running scalar spectral index*, $n_{run} \equiv dn_s/(d \ln k)$, which becomes

$$n_{run} = -2\left(9 - 5H\frac{\mathcal{F}_{,HH}}{\mathcal{F}_{,H}} + H^2\frac{\mathcal{F}_{,HHH}}{\mathcal{F}_{,H}}\right)\epsilon_H^2 + 2\left(8 - 3H\frac{\mathcal{F}_{,HH}}{\mathcal{F}_{,H}}\right)\epsilon_H\eta_H - 2\xi_H , \quad (31.46)$$

where ξ_H corresponds to the *third slow-roll parameter* and turns out to be given by

$$\xi_H \equiv \left(\frac{m_{Pl}^2}{4\pi}\right)^2 \left(\frac{\mathcal{F}_{,H}}{H}\right)^2 \frac{H'''H'}{H^2} . \quad (31.47)$$

In addition to the scalar curvature perturbation, transverse-traceless tensor perturbation can also be generated from quantum fluctuations during inflation [31.76, 80, 81]. The tensor perturbations do not couple to matter and, consequently, they are only determined by the dynamics of the background metric, so the standard results for the evolution of tensor perturbations of the metric remain valid. The two independent polarizations evolve like minimally coupled massless fields with spectrum (we mention here that this expression should be implemented with a factor such that $F_\alpha^2(H/\mu)$, where

$$F_\alpha^{-2}(x) = \sqrt{1+x^2} - \left(\frac{1-4\alpha\mu^2}{1+4\alpha\mu^2}\right)x^2 \sinh^{-1}\frac{1}{x} ,$$

when a brane-world with a Gauss–Bonnet term is considered [31.98].)

$$\mathcal{P}_{\mathcal{T}} = \frac{16\pi}{m_{Pl}^2}\left(\frac{H}{2\pi}\right)^2 \Bigg|_{aH=k} . \quad (31.48)$$

Similarly to the case of scalar perturbations, we evaluate the expression on the right-hand side of (31.48) when the comoving scale k leaves the horizon during inflation. Furthermore, we can introduce the *gravitational wave spectral index* n_T defined by $n_T \equiv (d \ln \mathcal{P}_{\mathcal{T}})/(d \ln k)$, which turns out to be

$$n_T = -2\epsilon_H . \quad (31.49)$$

Here, we can also introduce the *running tensor spectral index* α_T defined by

$$\alpha_T \equiv \frac{dn_T}{d \ln k} = 4\epsilon_H\eta_H - 2\left(3 - H\frac{\mathcal{F}_{,HH}}{\mathcal{F}_{,H}}\right)\epsilon_H^2 . \quad (31.50)$$

At this point we can define the *tensor-to-scalar ratio* $r \equiv \mathcal{P}_{\mathcal{T}}/\mathcal{P}_{\mathcal{R}}$, which becomes

$$r = 2\frac{\mathcal{F}_{,H}}{H}\epsilon_H , \quad (31.51)$$

and combining (31.49) and (31.51) we find that

$$n_T = -\left(\frac{H}{\mathcal{F}_{,H}}\right) r . \quad (31.52)$$

This latter expression corresponds to the *inflationary consistency condition* [31.92, 99]. Note that for standard cosmology, in which $\mathcal{F}(H)=H^2$, (31.52) reduces to $r=-(1/2)n_T$. However, this relation may be violated in some cases [31.100, 101]. Note that this relation depends on the kind of theory that we are dealing with.

31.5 Hierarchy of Slow-Roll Parameters and Flow Equations

There is a different way of studying inflationary universe models, which is subtended by a sort of hierarchy imposed on the slow-roll parameters [31.102, 103]. In fact, the set of equations in this approach is based on derivatives with respect to the *e-fold number* over the slow-roll parameters.

We previously introduced the slow-roll parameters, such as ϵ_H, η_H and ξ_H, to which we have given a sort of hierarchy, calling them *first*, *second*, and *third slow-roll parameters*, respectively. Each of these parameters is characterized by their dependence on the order of the scalar field derivative of the Hubble ratio $H(\phi)$, such as $\epsilon_H\approx(H')^2$, $\eta_H\approx H''$ and $\xi_H\approx H'''$, as we can see from (31.15), (31.35), and (31.47), respectively. It is possible to extend this definition to higher derivatives of the Hubble parameter so that we can introduce the following parameter

$$ {}^{l}\lambda_H\equiv\left(\frac{m_{Pl}^2}{4\pi}\right)^l\left(\frac{\mathcal{F}_{,H}}{H}\right)^l\frac{(H')^{l-1}}{H^l}\frac{d^{l+1}H}{d\phi^{l+1}}\,;\quad (l\geq 1)\,, \tag{31.53}$$

where for $l=1$ we have ${}^1\lambda_H\equiv\eta_H$ and $l=2$ corresponds to ${}^2\lambda_H\equiv\xi_H$.

It is not difficult to show that the following set of equations is satisfied

$$\begin{aligned}\frac{d\epsilon_H}{dN}&=\left[\left(H\frac{\mathcal{F}_{,HH}}{\mathcal{F}_{,H}}-3\right)\epsilon_H+2\eta_H\right]\epsilon_H\,,\\ \frac{d\,{}^l\lambda_H}{dN}&=\left[l\left(H\frac{\mathcal{F}_{,HH}}{\mathcal{F}_{,H}}-2\right)\epsilon_H+(l-1)\eta_H\right]{}^l\lambda_H\\ &\quad+{}^{l+1}\lambda_H\,;\quad (l\geq 1)\,.\end{aligned} \tag{31.54}$$

Here the relationship

$$\frac{d}{dN}\equiv-\frac{m_{Pl}^2}{8\pi}\frac{\mathcal{F}_{,H}}{H}\left(\frac{H'}{H}\right)\frac{d}{d\phi}$$

was used.

In the standard case, i.e., when $\mathcal{F}(H)=H^2$, the above set of equations reduces to

$$\begin{aligned}\frac{d\epsilon_H}{dN}&=\epsilon_H(\sigma+2\epsilon_H)\,,\\ \frac{d\sigma}{dN}&=-5\epsilon_H\sigma-12\epsilon_H^2+2\xi_H\,,\\ \frac{d\,{}^l\lambda_H}{dN}&=\left[\frac{l-1}{2}\sigma+(l-2)\epsilon_H\right]{}^l\lambda_H\\ &\quad+{}^{l+1}\lambda_H\,;\quad (l\geq 2)\,,\end{aligned} \tag{31.55}$$

where $\sigma\equiv 2\eta_H-4\epsilon_H$ [31.103].

In order to solve the infinite set of equations (31.55) the series is truncated by imposing a vanishing value to a given high enough slow-roll parameter. This corresponds to taking that ${}^{M+1}\lambda_H=0$, for an appropriate large number M (for instance, $M=5$ has been used in the literature [31.103]). With this truncation the set of equations has been solved both numerically [31.102–105] and analytically [31.106–108].

With respect to possible solutions of these equations and their relations with the inflationary paradigm, it was emphasized in [31.106] that there is no clear connection between them. Actually, it is not clear that a particular solution of the flow equations corresponds directly to some type of inflationary solution. To achieve this task we must add additional ingredients to the corresponding solutions. The main ingredient that has been left out of this scheme is the Friedmann equation itself. Thus, in solving the flow equations we were able to obtain ϵ_H as a function of the scalar field ϕ (imposing the condition that this parameter will satisfy the range $0\leq\epsilon_H\leq 1$), and then we were able to obtain $H(\phi)$ through the following relation

$$\int_{H_i}^{H(\phi)}\sqrt{\frac{\mathcal{F}_{,H}}{H^3}}\,dH=\sqrt{\frac{8\pi}{m_{Pl}^2}}\int_{\phi_i}^{\phi}\sqrt{\epsilon_H(\phi)}\,d\phi\,. \tag{31.56}$$

Thus, by obtaining an explicit expression for ϵ_H and giving an explicit expression for $\mathcal{F}(H)$ as a function of H, we obtained the Hubble parameter $H(\phi)$ as a function of the scalar field through (31.56) (in the standard

case in which $\mathcal{F} = H^2$ it is obtained that

$$H(\phi) = H_{\text{i}} \exp\left[\sqrt{\frac{8\pi}{m_{\text{Pl}}^2}} \int_{\phi_{\text{i}}}^{\phi} \sqrt{\epsilon_{\text{H}}(\phi)}\,\mathrm{d}\phi\right].$$

). With this in hand, we can obtain an explicit expression for the scalar field potential $V(\phi)$, given by

$$V(\phi) = \left(\frac{3m_{\text{Pl}}^2}{8\pi}\right)\mathcal{F}\left[1 - \frac{1}{6}H\left(\frac{\mathcal{F}_{,H}}{\mathcal{F}}\right)\epsilon_{\text{H}}\right]. \quad (31.57)$$

In short, for a given function $\mathcal{F}(H)$, we could say that the Hubble flow formalism allows us to determine the scalar field potential $V(\phi)$ associated to some inflationary universe model. In order to realize this task we first solve the flow equations, (31.54), from which we can obtain the first slow-roll parameter ϵ_{H} under the condition that this parameter must satisfy the bound $0 \le \epsilon_{\text{H}} \le 1$. Then, by using (31.56), we obtain the corresponding Hubble parameter as a function of the scalar field $H(\phi)$ from which we obtain all the other quantities associated to the inflationary scenario.

31.6 A Possible Way of Obtaining the Generating Function $H(\phi)$

There is a way to obtain the generating function, i. e., H as a function of the inflaton scalar field ϕ, explicitly. This approach was revealed for the first time in [31.109]. The procedure is as follows: we noticed that variable z, defined as $z = \frac{\dot{\phi}}{H}$, plays an important role in the description of scalar perturbations (see (31.37)). Actually, its second derivative with respect to the conformal time gives the *mass* term, (31.39).

On the other hand, with respect to inflationary universe models, we can restrict ourselves to the particular case in which the variable z is a constant at superhorizon scale [31.109]. By imposing this latter condition, i. e., $z =$ const., a straightforward calculation leads to the following differential equation for the Hubble parameter $H(\phi)$

$$\left(\frac{m_{\text{Pl}}^2}{8\pi}\right)\left[(H\mathcal{F}_{,HH} - 2\mathcal{F}_{,H})\left(\frac{H'}{H}\right)^2 + \left(\frac{\mathcal{F}_{,H}}{H^2}\right)H''\right] - H = 0\,. \quad (31.58)$$

In principle, after solving this latter differential equation we can obtain H as a function of the scalar field ϕ, and with this explicit expression for H we can obtain, for instance, the primordial curvature perturbations $\mathcal{P}_{\mathcal{R}}(k)$, together with the tensor perturbation, $\mathcal{P}_{\mathcal{T}}(k)$.

Let us take the unit in which $m_{\text{Pl}}^2/8\pi = 1$, and if we take the case in which $\mathcal{F}(H) = H^2$, we have that (31.58) simplifies to

$$\frac{1}{2} + \left(\frac{H'}{H}\right)^2 - \frac{H''}{H} = 0\,, \quad (31.59)$$

which presents a solution of the type [31.109]

$$H(\phi) = H_0 \exp\left(\frac{\phi^2}{4} + \phi_{\text{i}}\phi\right), \quad (31.60)$$

where H_0 and ϕ_{i} are two arbitrary constants. In this case, for instance, the scalar potential, the scale factor, and the cosmological time become a function of the scalar field given by (taking the constant $\phi_{\text{i}} = 0$)

$$V(\phi) = H_0^2\left(3 - \frac{\phi^2}{2}\right)\exp\left(\frac{\phi^2}{2}\right), \quad (31.61)$$

$$a(\phi) = \frac{\phi_0}{\phi} \quad (31.62)$$

and

$$t(\phi) = \frac{1}{2H_0}\left[\text{Ei}\left(-\frac{\phi_0^2}{4}\right) - \text{Ei}\left(-\frac{\phi^2}{4}\right)\right], \quad (31.63)$$

respectively. Here, ϕ_0 is an integration constant and Ei is the exponential integral function. One interesting thing in this specific case is that the equation governing the evolution of scalar perturbations simplifies and can be solved, and with this solution together with the assumption that the variable z remains constant, it is possible to calculate the spectral index, which turns out to be exact and ϕ-independent, namely,

$$n_{\text{s}} = 3\,.$$

Unfortunately, this result, when compared with that corresponding to an observed scale-free spectrum (which is close to unity) presents a large blue shift. Apart from this, we could say that this case and the one related to the Einstein power-law inflation are the only ones that have this feature [31.109].

31.7 Two Interesting Cases

In the following descriptions we will study the two cases that we mentioned above, in the introductory section, i. e., the Friedmann–Chern–Simons and the brane-world type of inflationary universe models.

31.7.1 The Friedmann–Chern–Simons Model

As stated in the introductory section, we would like to consider here a model in which the Friedmann equation modifies to

$$\mathcal{F}(H) \equiv H^2 - \alpha H^4 = \left(\frac{8\pi}{3m_{\rm Pl}^2}\right)\rho_\phi\,, \tag{31.64}$$

where α is an arbitrary constant with dimension $m_{\rm Pl}^{-2}$. Here, we assume that during the inflationary evolution the Hubble parameter H satisfies the bound $H < 1/\sqrt{\alpha}$, so that the energy density associated to the scalar field ϕ is positive.

For the generating function it is possible to choose a polynomial like $H(\phi) = H_0(1 + \beta\phi + \beta_2\phi^2 + \ldots\beta_N\phi^N)$, where H_0 and the different β are constants. This sort of solution was used to generate suitable functions of slow-roll parameters [31.106]. Here, just for simplicity, and in order to show how this approach works, we shall take the previous polynomial, but, up to first order in the scalar field ϕ, i. e., $H(\phi) = H_0(1 + \beta\phi)$, with β an arbitrary constant with dimension $m_{\rm Pl}^{-1}$. In this case, the scalar potential becomes

$$V(\phi) = \left(\frac{3m_{\rm Pl}^2}{8\pi}\right) H_0^2\bar{\phi}^2\left[1 - \alpha H_0^2\bar{\phi}^2\right] \times\left[1 - \frac{m_{\rm Pl}^2}{12\pi}\frac{\beta^2}{\bar{\phi}^2}\left(\frac{1 - 2\alpha H_0^2\bar{\phi}^2}{\sqrt{1 - \alpha H_0^2\bar{\phi}^2}}\right)^2\right], \tag{31.65}$$

where $\bar{\phi} \equiv 1 + \beta\phi$.

In the slow-roll approximation, i. e., where $\dot{\phi}^2 \ll V(\phi)$ together with $|\ddot{\phi}| \ll |\mathrm{d}V(\phi)/\mathrm{d}\phi|$, it is found that the scalar field potential becomes

$$V(\phi)_{s-r} \simeq \left(\frac{3m_{\rm Pl}^2}{8\pi}\right) H_0^2\bar{\phi}^2\left(1 - \alpha H_0^2\bar{\phi}^2\right). \tag{31.66}$$

Figure 31.7 depicts the shape of the potential for the exact case (thick line), expressed by (31.65), together with the approximated slow-roll case, (31.66). In the figure the dotted line represents the exact case in which the α-parameter vanishes.

Here, from expression (31.15) we have that in this case the first Hubble slow-roll parameter $\epsilon_{\rm H}$ is given by

$$\epsilon_{\rm H} \equiv -\frac{\mathrm{d}\ln H}{\mathrm{d}\ln a} = \left(\frac{m_{\rm Pl}^2}{4\pi}\right)\left(1 - 2\alpha H^2\right)\left(\frac{H'}{H}\right)^2. \tag{31.67}$$

This parameter gives information about the acceleration of the Universe. During inflation we have $\epsilon_{\rm H} < 1$, and it ends when $\epsilon_{\rm H}$ takes the value equal to 1.

The amount of inflation is

$$N(t) \equiv \ln\frac{a(t_{\rm end})}{a(t)}\,, \tag{31.68}$$

where $a(t_{\rm end})$ is the scale factor evaluated at the end of inflation. Thus, we have

$$N(\phi) = \int_t^{t_{\rm end}} H\,\mathrm{d}t = \int_{\phi_{\rm end}}^{\phi} \frac{1}{\epsilon_{\rm H}}\frac{H'}{H}\,\mathrm{d}\phi\,. \tag{31.69}$$

Here, $\phi_{\rm end}$ represents the value of the scalar field at the end of inflation. Its value is determined by imposing $\epsilon_{\rm H}(\phi_{\rm end}) = 1$.

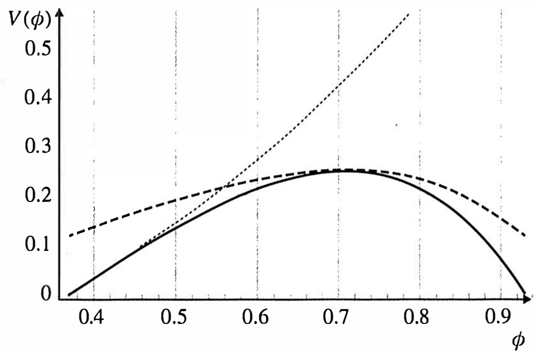


Fig. 31.7 Plots of the scalar potentials $V(\Phi)$ as a function of the dimensionless *scalar field*, $\Phi \equiv \sqrt{\alpha}H_0\bar{\phi}$. The *thick line* represents the exact potential, expressed by (31.65). The *dashed line* represents the same potential, but in the slow-roll approximation, (31.66). The *dotted line* corresponds to the exact case, but when $\alpha = 0$. Here we have taken $\alpha(\beta H_0)^2 \equiv (24\pi)/(9m_{\rm Pl}^2)$ and $V(\Phi)$ is expressed as a multiple of the constant $V_0 \equiv (3m_{\rm Pl}^2)/(8\pi\alpha)$

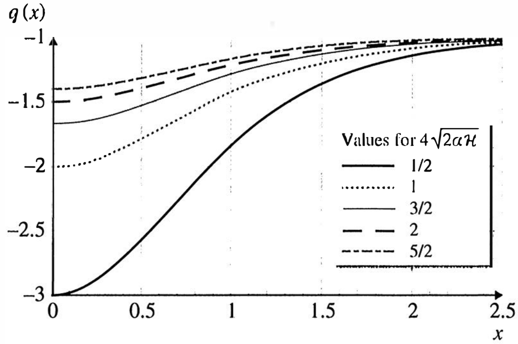


Fig. 31.8 Time evolution of the deceleration parameter q as a function of $x = \mathcal{H}(t - t_0)$. Here we have taken the value $\sqrt{\alpha}H_0 = 19/(5\sqrt{2})$. The *thick*, *dotted*, *thin*, *dashed*, and *dot-dashed lines* correspond to the values for $4\sqrt{2\alpha}\mathcal{H} = 1/2; 1; 3/2; 2; 5/2$, respectively. Note that in all cases the acceleration parameters turn out to be negative

The scalar field results are given by

$$\phi(t) = -\frac{1}{\beta} + \frac{1}{\sqrt{2\alpha}\beta H_0} \times \cosh\left[2\tanh^{-1}\left(\tanh\left[\frac{1}{2}\cosh^{-1}(\sqrt{2\alpha}H_0)\right] \times e^{\mathcal{H}(t-t_0)}\right)\right], \tag{31.70}$$

where $\mathcal{H} \equiv \sqrt{2\alpha}(\beta H_0)^2 \frac{m_{\rm Pl}^2}{4\pi}$ and $\phi(t_0) = 0$. This latter expression allows us to write down the Hubble parameter as a function of time. From this result we obtain the scale factor $a(t)$, which becomes

$$a(t) = a_0\left(\frac{\sinh\{2\tanh^{-1}\left[\tanh(\frac{1}{2}\cosh^{-1}(\sqrt{2\alpha}H_0)) \times e^{\mathcal{H}(t-t_0)}\right]\}}{\sinh\left[\cosh^{-1}(\sqrt{2\alpha}H_0)\right]}\right)^{4\sqrt{2\alpha}\mathcal{H}}. \tag{31.71}$$

In order to see if this latter expression describes an accelerated phase, for given values of the parameters, in Fig. 31.8 we plot the deceleration parameter q, which is defined as $q = -(\ddot{a}a)/\dot{a}^2$. For this plot we have taken the value $\sqrt{\alpha}H_0 = \frac{19}{5\sqrt{2}}$. The different curves correspond to different values of the exponent that appears in the scale factor a, i. e., $4\sqrt{2\alpha}\mathcal{H}$. These curves show that the universe is accelerating, since the parameter $q(t)$ turns out to be negative as time passes. Therefore, our model presents a period of inflation, at least for the values of the parameters that we have considered here.

The amount of inflation in this case becomes

$$\overline{N}(y) = -\overline{N}_{\rm e} + \sqrt{\gamma} \times \left[\frac{1}{\sqrt{y}} + \frac{1}{\sqrt{2}}\left(\frac{1-4\gamma}{\gamma}\right)\tanh^{-1}\left(\sqrt{2y}\right)\right], \tag{31.72}$$

where y is a dimensionless function of the scalar field defined by $y = \alpha H_0^2(1 + \beta\phi)^2$, γ is a dimensionless constant given by $\gamma \equiv (m_{\rm Pl}^2/4\pi)\alpha(H_0\beta)^2$, and $\overline{N}_{\rm e}$ corresponds to

$$\overline{N}_{\rm e} = \frac{1}{2}\sqrt{1+2\gamma} + \frac{1}{2\sqrt{2}}\left(\frac{1-4\gamma}{\sqrt{\gamma}}\right)\tanh^{-1}\left(\sqrt{\frac{2\gamma}{1+2\gamma}}\right). \tag{31.73}$$

Let us now consider the attractor behavior of this model. By taking into account (31.21), we have

$$\delta H(\phi) = \delta H(\phi_{\rm i})\exp\left[\frac{12\pi}{m_{\rm Pl}^2}\int_{\phi_{\rm i}}^{\phi} g(H_0)\frac{H_0}{H_0'}\,{\rm d}\phi\right], \tag{31.74}$$

where $\phi_{\rm i}$ represents the initial value of the scalar field ϕ. The function $g(H_0)$ is given by

$$\frac{\left[1 - 2\alpha H_0^2\left(1 - \frac{2}{3}\epsilon_{H_0}\alpha H_0^2\right)\right]}{\left(1 - 2\alpha H_0^2\right)^2},$$

and it is positive for $2\alpha H_0^2 < 1$ (this makes sure that the energy density will be positive, as we can see from (31.64)). Thus, the integrand within the exponential term will be negative, due to the fact that ${\rm d}\phi$ and H_0' have contrary signs (assuming that the perturbation δH does not change the sign of $\dot{\phi}$)[31.61]. In this way, all the linear perturbations tend to vanish rapidly.

Regarding scalar perturbations, the scalar spectral index parameter becomes

$$n_{\rm s} - 1 = 2\eta_{\rm H} - 4\left(1 - \frac{2\alpha H^2}{1 - 2\alpha H^2}\right)\epsilon_{\rm H}, \tag{31.75}$$

and the running scalar spectral index is

$$n_{\text{run}} = \left(\frac{10}{1-2\alpha H^2}\right)\epsilon_{\text{H}}\eta_{\text{H}} + \left(\frac{8}{1-2\alpha H^2}\right)\epsilon_{\text{H}}^2 - 2\xi_{\text{H}}^2 , \tag{31.76}$$

where ξ_{H} is defined as

$$\xi_{\text{H}}^2 \equiv \left(\frac{m_{\text{Pl}}^2}{4\pi}\right)^2 (1-2\alpha H^2)^2 \frac{H'''H'}{H^2} . \tag{31.77}$$

Analogously, from (31.49) we found an expression for the gravitational wave spectral index n_{T}, given by

$$n_{\text{T}} = -2\epsilon_{\text{H}} , \tag{31.78}$$

and the corresponding tensor-to-scalar amplitude ratio

$$r = 4\left(1-2\alpha H^2\right)\epsilon_{\text{H}} . \tag{31.79}$$

Bearing in mind the expression that we have required for $H(\phi)$, we obtain a relationship between r and n_s given by

$$r(n_s) = \frac{(8\gamma + 1 - n_s)^2}{16\gamma + 1 - n_s} , \tag{31.80}$$

where the dimensionless constant γ was defined previously.

Note that we need to satisfy $n_s < 1 + 16\gamma$ in order to have $r > 0$. Thus, from this inequality we obtain a constraint on the parameter γ given by $\gamma > \frac{1}{16}|(n_s - 1)|$.

Figure 31.9 shows how r changes as a function of n_s for two different values of the parameter γ. These values are $\gamma = 8.0 \times 10^{-3}$ and $\gamma = 16.0 \times 10^{-3}$. From this figure we see that our model can accommodate the observational data quite well. Note that this model allows the possibility of having a Harrison–Zel'dovich spectrum, i. e., $n_s = 1$, with $r \neq 0$ as could be seen from the plot.

In this case, the system of flow equations (31.54) is reduced to the following set of equations

$$\frac{d\epsilon_{\text{H}}}{dN} = \left[2\left(\frac{1-4\alpha H^2}{1-2\alpha H^2}\right)\epsilon_{\text{H}} + \sigma\right]\epsilon_{\text{H}} ,$$

$$\frac{d\sigma}{dN} = -\left[6\left(\frac{2-\alpha H^2}{1-2\alpha H^2}\right)\epsilon_{\text{H}} + \left(\frac{5-6\alpha H^2}{1-2\alpha H^2}\right)\sigma\right]\epsilon_{\text{H}} + 2\xi_{\text{H}} ,$$

$$\frac{d\,{}^l\lambda_{\text{H}}}{dN} = \left[l\left(\frac{1-6\alpha H^2}{1-2\alpha H^2}\right)\epsilon_{\text{H}} + \frac{1}{2}(l-1)\sigma\right] \times {}^l\lambda_{\text{H}} + {}^{l+1}\lambda_{\text{H}} \quad (l \geq 2) . \tag{31.81}$$

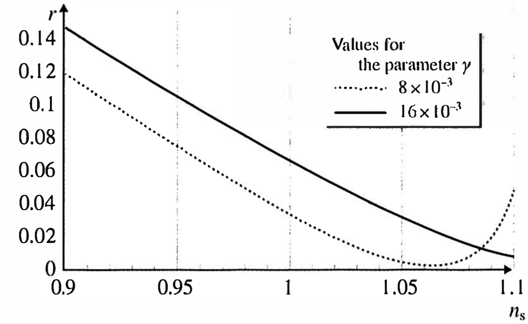


Fig. 31.9 The parameter r as a function of the scalar spectral index n_s for two values of the constant $\gamma = (m_{\text{Pl}}^2/4\pi)\alpha\,(H_0\beta)^2$, as described by (31.80). Here, we have taken the values $\gamma = 8.0\times10^{-3}$ and $\gamma = 16.0\times10^{-3}$. Note that we could have the possibility of having a Harrison–Zel'dovich spectrum ($n_s = 1$) with $r \neq 0$.

In order to solve this set of equations we need to have $H = H(N)$. In order to obtain this result, we start by considering (31.16), which results $N = N(\phi)$. Then, we need to invert this latter expression (if possible) to obtain $\phi = \phi(N)$. Finally, with this expression we obtain H as a function of N, and, by introducing this function into the flow equation, it is possible to solve it.

There is another way of obtaining a relationship between H and N. Let us assume that we really know the Hubble rate as a function of the scale factor, i. e., we know $H(a)$ explicitly. Then, since by definition $dN = -d\ln a$, we obtain $a(N) = a_e e^{(N_e - N)}$, where a_e and N_e are the values of the scale factor and the number of e-folds at the end of inflation. Then, by a direct substitution of $a(N)$ on the Hubble rate H we obtain $H(N)$.

As an example of the latter approach, let us consider the model in which there is a smooth exit from inflation, under the so-called *decaying vacuum cosmology* [31.110]. There, it was found that the Hubble parameter as a function of the scale factor becomes

$$H(a) = 2H_e\left(\frac{a_e^2}{a^2 + a_e^2}\right) , \tag{31.82}$$

where $H_e = H(a_e)$. In this case it is obtained that

$$H(N) = H_e\left[1 - \tanh\left(N_e - N\right)\right] . \tag{31.83}$$

Let us solve numerically the set of (31.81) for the first two slow-roll parameters, ϵ_{H} and η_{H}, when ξ_{H} is a constant equal to 0.2. Figure 31.10 shows the numerical solutions for these two slow-roll parameters. From the figure we can see that ϵ_{H} remains almost

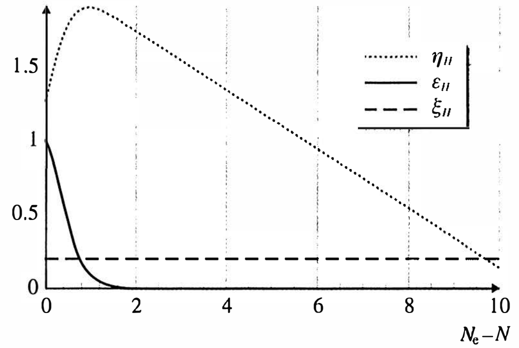


Fig. 31.10 Numerical solutions for $\epsilon_{\rm H}$ and $\eta_{\rm H}$ from the set of equations (31.81) in the case in which $\xi = \text{const.} = 0.2$ and $H(N) = H_e[1 - \tanh(N_c - N)]$

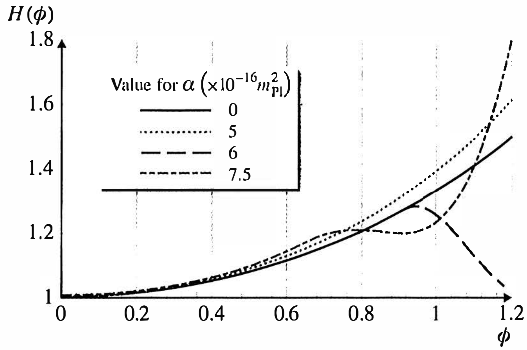


Fig. 31.11 The evolution of the Hubble parameter $H(\phi)$ as a function of the scalar field ϕ. Here we have plotted $H(\phi)$ for the values $\alpha(\times 10^{-16} m_{\rm Pl}^2) = 0.0$; 5.0; 6.0, and 7.5

constant (closed to zero) for a wide range of values of $\widetilde{N} \equiv N_e - N$. However, for $\widetilde{N} < 1$ it increases to the value of 1. Actually, for $N = N_e$, i. e., at the end of inflation, $\epsilon_{\rm H} = 1$. In the same range, i. e., $\widetilde{N} < 1$, the other slow-roll parameter $\eta_{\rm H}$ decreases from a maximum value (closed to the point $\widetilde{N} \approx 1$) to its final value $\eta_{\rm H} \approx 1.2$ at the end of inflation. For this parameter, in the case $\widetilde{N} > 1$, is can be observed from the figure that it decreases lineally.

Applying the approach followed in [31.109] in this case, we obtain that the differential (31.58) reduces to

$$\frac{H''}{H} - \left(\frac{1+2\alpha H^2}{1-2\alpha H^2}\right)\left(\frac{H'}{H}\right)^2 - \left(\tfrac{1}{2}\right)\left(\frac{1}{1-2\alpha H^2}\right) = 0\,. \tag{31.84}$$

This ordinary differential equation is quite hard to solve analytically. Figure 31.11 presents numerical solutions for different values of α. It shows how the Hubble parameter $H(\phi)$ changes as a function of the scalar field ϕ. For comparison we have included the exact solution corresponding to the $\alpha = 0$ case.

31.7.2 The Brane-World Model

As we mentioned in Sect. 31.1, we consider a five-dimensional brane scenario in which the Friedmann equation is modified to [31.51–53]

$$H^2 = \left(\frac{8\pi}{3m_{\rm Pl}^2}\right)\rho_\phi\left(1 + \frac{\rho_\phi}{2\lambda}\right), \tag{31.85}$$

where λ represents the brane tension.

The previous expression was also considered in the high energy regime [31.111], i. e., when $\rho_\phi/2\lambda \gg 1$. In this case the function $\mathcal{F}(H)$ is

$$\mathcal{F}(H) = \sqrt{\frac{16\pi\lambda}{3m_{\rm Pl}^2}}H\,.$$

With this expression for the function $\mathcal{F}(H)$, for the scalar field potential we obtain

$$V(\phi) = \sqrt{\frac{3\lambda m_{\rm Pl}^2}{4\pi}}H(\phi) - \frac{\lambda m_{\rm Pl}^2}{24\pi}\left(\frac{H'(\phi)}{H(\phi)}\right)^2.$$

Here we assume that the Hubble factor presents an exponential dependence, i. e., $H(\phi) \approx \exp(-\phi)$.

Expression (31.85) can be written as

$$\mathcal{F}(H) \equiv b\left[\sqrt{1+\left(\frac{2}{b}\right)H^2} - 1\right] = \left(\frac{8\pi}{3m_{\rm Pl}^2}\right)\rho_\phi\,, \tag{31.86}$$

where b is defined by $b \equiv (8\pi\lambda)/(3m_{\rm Pl}^2)$.

From this latter equation, together with (31.11), we obtain for the scalar potential

$$V(\phi) = \lambda\left[\sqrt{1+\left(\frac{2}{b}\right)H^2} - 1\right] - \tfrac{2}{9}\left(\frac{\lambda}{b^2}\right)\frac{(H')^2}{\left[1+\left(\frac{2}{b}\right)H^2\right]}\,. \tag{31.87}$$

In order to obtain an explicit expression for the scalar potential we need to introduce an explicit expression for the Hubble parameter as a function of the

scalar field. In this respect, we borrow the expression put forward by *Hawkins* and *Lidsey* for the Hubble parameter [31.30]. Thus, we take

$$H(\phi) = \sqrt{\frac{b}{2}}\left(\frac{\coth(\beta\phi)}{\sinh(\beta\phi)}\right),$$

where β is a constant given by $\beta \equiv (\sqrt{2\pi}C)/(m_{\mathrm{Pl}})$, with C an arbitrary dimensionless constant.

The scalar field potential and the scale factor become

$$V(\phi) = \frac{\lambda}{3}(6-C^2)\mathrm{csch}^2\left(\frac{\sqrt{2\pi}C}{m_{\mathrm{Pl}}}\phi\right) \tag{31.88}$$

and

$$a(t) = \frac{1}{2}bC^4\left[\left(t+\frac{4}{C^2\sqrt{b}}\right)t\right]^{1/C^2}, \tag{31.89}$$

respectively [31.30]. Two comments are in order, first we demand that C be less than $\sqrt{6}$ in order for the potential to be positive definite, and secondly, in the expression for the scale factor we have chosen $t_0 = -((3m_{\mathrm{Pl}}^2)/(4\pi\lambda C^4))^{1/2}$ in order to have $a(0) = 0$. For an early time it is found that $a \approx t^{1/C^2}$, therefore, for inflation to be realizable we need $C^2 < 1$.

In this case, the amount of comoving inflation becomes

$$\overline{N}(x) = \ln\left[\left(\frac{\sinh(x_{\mathrm{e}})}{\sinh(x)}\right)^{\frac{2}{C^2}(1-C^2)}\frac{\cosh(x_{\mathrm{end}})}{\cosh(x)}\right]. \tag{31.90}$$

Here, $x \equiv (\sqrt{2\pi}C)/(m_{\mathrm{Pl}})\phi$ and $x_{\mathrm{e}} = (\sqrt{2\pi}C)/(m_{\mathrm{Pl}})\phi_{\mathrm{e}}$, where ϕ_{e} is the value of the scalar field at the end of inflation, which corresponds to

$$\phi_{\mathrm{e}} = \frac{m_{\mathrm{Pl}}}{\sqrt{2\pi}C}\mathrm{sech}^{-1}\left[\frac{1}{C}\sqrt{2-C^2}\right].$$

Regarding the attractor solution, from (31.21) we obtain

$$\delta H(\phi) = \delta H(\phi_{\mathrm{i}}) \times \exp\int_{\phi_{\mathrm{i}}}^{\phi}\left(\frac{3}{\epsilon_{\mathrm{H}}}+\frac{2}{b}\frac{H^2}{1+\frac{2}{b}H^2}\right)\left(\frac{H'}{H}\right)\Bigg|_0 \mathrm{d}\phi. \tag{31.91}$$

The quantity in square brackets is positive definite, thus the difference in sign between H' and $\mathrm{d}\phi$ makes the exponential negative, and therefore, the exponential rapidly tends to zero, showing the attractor feature.

Returning to the previously introduced expression for the Hubble parameters, for the various slow-roll parameters we obtain the following expressions

$$\epsilon_{\mathrm{H}}(\phi) = \frac{C^2}{2}\left[1+\mathrm{sech}^2(\beta\phi)\right], \tag{31.92}$$

$$\eta_{\mathrm{H}}(\phi) = \frac{C^2}{2}\left(1+\frac{8}{3+\cosh(2\beta\phi)}\right) \tag{31.93}$$

and

$$\xi_{\mathrm{H}} = \frac{C^4}{4}\left[1+5\mathrm{sech}^2(\beta\phi)+\frac{24}{3+\cosh(2\beta\phi)}\right]. \tag{31.94}$$

By using these expressions we obtain n_{s} and r, which become

$$n_{\mathrm{s}} - 1 = -\frac{C^2}{2}\mathrm{sech}^2(\beta\phi)\left[5+\cosh(2\beta\phi)\right]$$

and

$$r = 2C^2\tanh(\beta\phi),$$

respectively. It is not difficult to show that the following relation holds

$$r = 3C^2 + (n_{\mathrm{s}} - 1). \tag{31.95}$$

Now, due to the observational constraint on r, which presents an upper limit, $r < 0.20$ (95% CL) from WMAP + BAO + SN [31.112], where SN is the constitution samples compiled in [31.113], and since $n_{\mathrm{s}} = 0.963 \pm 0.012$ (excluding the Harrison–Zel'dovich spectrum in a value greater than 3σ) [31.114], we have that the parameter C should satisfy the upper bound $C^2 < 0.079 \pm 0.004$ in order to be in agreement with the observational data.

On the other hand, the consistency condition in this case becomes

$$r = -\frac{2}{\sqrt{1+\frac{2}{b}H^2}}n_{\mathrm{T}},$$

where n_{T} turns out to be $n_{\mathrm{T}} = -C^2\left[1+\mathrm{sech}(\beta\phi)\right]$. Note that in the limit in which $b \longrightarrow \infty$ we obtain the standard results $r = -2n_{\mathrm{T}}$. Concerning the hierarchy

slow-roll parameters equations we find

$$\frac{d\epsilon_H}{dN} = \left[2\left(\frac{1+\frac{1}{b}H^2}{1+\frac{2}{b}H^2}\right)\epsilon_H + \sigma\right]\epsilon_H ,$$
$$\frac{d\sigma}{dN} = +2\xi_H - \left[\left(\frac{5+\frac{12}{b}H^2}{1+\frac{2}{b}H^2}\right)\sigma + 12\epsilon_H\right]\epsilon_H ,$$
$$\frac{d^l\lambda_H}{dN} = \left[\left(\frac{l-2-\frac{4}{b}H^2}{1+\frac{2}{b}H^2}\right)\epsilon_H + \frac{1}{2}(l-1)\sigma\right] \times {}^l\lambda_H + {}^{l+1}\lambda_H \qquad (l \geq 2) . \tag{31.96}$$

Following an approach analogous to the previous subsection we solve this set numerically in the case in which the ξ_H parameter remains constant equal to 0.2, and we use expression (31.83) for the dependence of the Hubble parameter as a function of the number of e-folds. The result is shown in Fig. 31.12. Note that η_H increases enormously close to the end of inflation. With this parameter much greater than 1 and ϵ_H reaching the value equal to 1 at the end of inflation, the slow-roll approximation becomes unsustainable at the end of inflation.

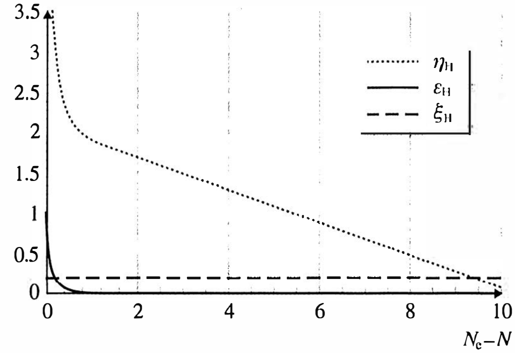


Fig. 31.12 Numerical solutions for ϵ_H and η_H from the set of equations (31.96) in the case in which $\xi = const. = 0.2$. Here, $H(N) = H_e\,[1 - \tanh(N_e - N)]$ was used

31.8 Conclusion

In this chapter we have given a general description of inflationary universe models in the case of modified Friedmann equations of the type $\mathcal{F}(H) \equiv (8\pi)/(3m_{Pl}^2)\rho_\phi$ within the scheme referred to as the exact Hamilton–Jacobi approach.

First, we introduced different types of inflationary universe models, which were classified in terms of small, large, and hybrid types of models.

Then, we gave a description of the exact approach to inflationary universe models. We introduced different definitions and a study of attractor solutions. Some examples were described under this approach.

We studied inflationary universe models in terms of a single scalar field. We applied the exact solution approach to the modified Friedmann equations. After describing the main characteristics of the inflationary model in general terms, we described some details of two specific models. First, we studied a model characterized by a modified Friedmann equation of the type $H^2 - \alpha H^4 = (3m_{Pl}^2/8\pi)\rho_\phi$, where the kinematical evolution was described for the case in which the Hubble parameter evolves as $H(\phi) = H_0(1+\beta\phi)$. With this at hand, we obtained the scalar potential, the corresponding number of e-folds and the attractor feature of the model. For some values of the parameters that entered into the scenario, we were able to characterize inflationary universe models.

Concerning scalar and tensor perturbations we calculated the scalar and tensor power spectrum generated by the quantum fluctuations of the scalar and the gravitational fields. We determined scalar and tensor spectrum indices in terms of the so-called slow-roll parameters ϵ_H, η_H, and ξ_H. From these quantities we were able to write down explicit expressions for the different parameters. Moreover, the shape of the contours in the $r - n_s$ plane resulted in being in agreement with those given by WMAP 7. In fact, we found that the tensor-to-scalar ratio can adequately accommodate the currently available observational data for some values of the parameters.

In the case of the brane-world model, the functional form for the Hubble parameter was considered

$$H(\phi) = \sqrt{\frac{b}{2}}\left(\frac{\coth(\beta\phi)}{\sinh(\beta\phi)}\right) ,$$

where $\beta = (\sqrt{2\pi}C)/m_{Pl}$ is constant. With this expression we were able to determine all the kinematics and

dynamics of the model. On the other hand, current astrophysical data put an upper bound on the constant C, which becomes $C^2 < 0.079 \pm 0.004$.

An important point that we did not consider here was the reheating period. In general terms, inflation is a period of supercooled expansion such that, when inflation ends, the temperature of the universe needs to go up to a value that coincides with that corresponding to the temperature of the radiation epoch, which thus matches the big bang model. This issue, as far as we know, has not been studied under the exact approach. Perhaps this study may give some insight on a deeper understanding of the period of reheating.

On the other hand, and not least, we have considered only one scalar field in our approach. It would be interesting to develop this approach, i. e., the exact approach, out of the slow-roll approximation, where two or more fields enter into the inflationary picture. This would be interesting for the mere fact that non-Gaussianity is a point that deserves to be considered. By non-Gaussianity we mean the non-Gaussian contributions to the correlations of cosmological fluctuations that became important probes of the early Universe. In particular, it will play an important role in our understanding of fundamental aspects of cosmology, especially in understanding the physics of the very early Universe that created the primordial seeds for large-scale structures with the subsequent growth of structures via gravitational instability. Actually, we should mention that a large non-Gaussianity can be generated under the mechanism of single field inflation, but it is necessary to include a noncanonical kinetic term [31.115].

In the scheme of a single inflationary universe field within the Hamilton–Jacobi approach, all these points deserve to be considered in depth in further studies. We hope to address these points in the near future.

References

31.1 P.J.E. Peebles, D.N. Schramm, E.L. Turner, R.G. Kron: The case for the relativistic hot big bang cosmology, Nature **352**, 769 (1991)

31.2 P.J.E. Peebles: *Principles of Physical Cosmology* (Princeton Univ. Press, Princeton 1993)

31.3 P.J.E. Peebles, D.N. Schramm, E.L. Turner, R.G. Kron: The evolution of the universe, Sci. Am. **271**, 29 (1994)

31.4 S. Weinberg: *Cosmology* (Oxford Univ. Press, New York 2008)

31.5 A.A. Penzias, R.W. Wilson: A measurement of excess antenna temperature at 4080 Mc/s, Astrophys. J. **142**, 419 (1965)

31.6 G. Gamow: Expanding universe and the origin of elements, Phys. Rev. **70**, 572 (1946)

31.7 R.A. Alpher, H. Bethe, G. Gamow: The origin of chemical elements, Phys. Rev. **73**, 803 (1948)

31.8 R.A. Alpher, H. Herman: Reflections on early work on 'big bang' cosmology, Phys. Today **41**, 24 (1988)

31.9 E. Hubble: A relation between distance and radial velocity among extra-galactic nebulae, Proc. Natl. Acad. Sci. **15**, 168 (1929)

31.10 E. Hubble, M. Humason: The velocity-distance relation among extra-galactic nebulae, Astrophys. J. **74**, 43 (1931)

31.11 A. Riess, A.G. Riess, A.V. Filippenko, P. Challis, A. Clocchiatti, A. Diercks, P.M. Garnavich, R.L. Gilliland, C.J. Hogan, S. Jha, R.P. Kirshner, B. Leibundgut, M.M. Phillips, D. Reiss, B.P. Schmidt, R.A. Schommer, R.C. Smith, J. Spyromilio, C. Stubbs, N.B. Suntzeff, J. Tonry: Observational evidence from supernovae for an accelerating universe and a cosmological constant, Astronom. J. **116**, 1009 (1998)

31.12 S. Perlmutter, G. Aldering, G. Goldhaber, R.A. Knop, P. Nugent, P.G. Castro, S. Deustua, S. Fabbro, A. Goobar, D.E. Groom, I.M. Hook, A.G. Kim, M.Y. Kim, J.C. Lee, N.J. Nunes, R. Pain, C.R. Pennypacker, R. Quimby, C. Lidman, R.S. Ellis, M. Irwin, R.G. McMahon, P. Ruiz-Lapuente, N. Walton, B. Schaefer, B.J. Boyle, A.V. Filippenko, T. Matheson, A.S. Fruchter, N. Panagia, H.J.M. Newberg, W.J. Couch, The Supernova Cosmology Project: Measurements of Ω and Λ from 42 High-Redshift Supernovae, Astrophys. J. **517**, 565 (1999)

31.13 A. Guth: Inflationary universe: A possible solution to the horizon and flatness problems, Phys. Rev. D **23**, 347 (1981)

31.14 A.D. Linde: *Particle Physics and Inflationary Cosmology* (Gordon and Breach, New York 1990)

31.15 WMAP Collaboration: First year Wilkinson microwave anisotropy probe (WMAP) observations: Implications for inflation, Astrophys. J. Suppl. **148**, 213 (2003)

31.16 J.D. Barrow: Varieties of expanding universe, Class. Quantum Gravity **13**, 2965 (1996)

31.17 J.D. Barrow, N.J. Nunes: Dynamics of "logamediate" inflation, Phys. Rev. D **76**, 043501 (2007)

31.18 P.G. Ferreira, M. Joyce: Cosmology with a primordial scaling field, Phys. Rev. D **58**, 023503 (1998)

31.19 P. Binetruy: Models of dynamical supersymmetry breaking and quintessence, Phys. Rev. D **60**, 063502 (1999)

31.20 D.H. Lyth, A. Riotto: Particle physics models of inflation and the cosmological density perturbation, Phys. Rep. **314**, 1 (1999)

31.21 E.W. Kolb, M.S. Turner: *The Early Universe* (Addison-Wesley, Menlo Park 1990)
31.22 L. Kofman, A.D. Linde: Problems with tachyon inflation, J. High Energy Phys. **02**, 004 (2002)
31.23 G. Felder, L. Kofmann, A.D. Linde: Inflation and preheating in nonoscillatory models, Phys. Rev. D **60**, 103505 (1999)
31.24 B. Feng, M. Li: Curvaton reheating in non-oscillatory inflationary models, Phys. Lett. B **564**, 169 (2003)
31.25 M. Dime, L. Randall, S. Thomas: Supersymmetry breaking in the early universe, Phys. Rev. Lett. **75**, 398 (1995)
31.26 D.H. Lyth, D. Wands: Generating the curvature perturbation without an inflation, Phys. Lett. B **524**, 5 (2002)
31.27 S. del Campo, R. Herrera, J. Saavedra, C. Campuzano, E. Rojas: Curvaton reheating in a logamediate inflationary model, Phys. Rev. D **80**, 123531 (2009)
31.28 B.J. Carr, J.E. Lidsey: Primordial black holes and generalized constraints on chaotic inflation, Phys. Rev. D **48**, 543 (1993)
31.29 J.E. Lidsey: Towards a solution of the Omega-problem in power law and chaotic inflation, Class. Quantum Gravity **8**, 923 (1991)
31.30 R.M. Hawkins, J.E. Lidsey: Inflation on a single brane: Exact solutions, Phys. Rev. **63**, 041301 (2001)
31.31 L.P. Grishchuk, Y.V. Sidorav: Boundary conditions and the wave function of the universe. In: *Fourth Seminar on Quantum Gravity*, ed. by M.A. Markov, V.A. Berezin, V.P. Frolov (World Scientific, Singapore 1988)
31.32 A.G. Muslinov: On the scalar field dynamics in a spatially flat Friedman universe, Class. Quantum Gravity **7**, 231 (1990)
31.33 D.S. Salopek, J.R. Bond, J.M. Bardeen: Designing density fluctuation spectra in inflation, Phys. Rev. D **40**, 1753 (1989)
31.34 J.E. Lidsey, A.R. Liddle, E.W. Kolb, E.J. Copeland, T. Barreiro, M. Abney: Reconstructing the inflaton potential – An overview, Rev. Mod. Phys. **69**, 373 (1997)
31.35 S. del Campo: Exact solutions in inflation, AIP Conf. Proc. **1471**, 27 (2011)
31.36 J.D. Barrow: Graduated inflationary universes, Phys. Lett. B **235**, 40 (1990)
31.37 J.D. Barrow, P. Saich: The behaviour of intermediate inflationary universes, Phys. Lett. B **249**, 406 (1990)
31.38 A. Vallinotto, E.J. Copeland, E.W. Kolb, A.R. Liddle, D.A. Steer: Inflationary potentials yielding constant scalar perturbation spectral indices, Phys. Rev. D **69**, 103519 (2004)
31.39 A.A. Starobinsky: Inflaton field potential producing the exactly flat spectrum of adiabatic perturbations, J. Exp. Theor. Phys. Lett. **82**, 169 (2005)
31.40 A.A. Starobinsky: Inflaton field potential producing the exactly flat spectrum of adiabatic perturbations, Pisma Zhurnal Eksp. Teor. Fiz. **82**, 187 (2005)
31.41 J.D. Barrow, A.R. Liddle, C. Pahud: Intermediate inflation in light of the three-year WMAP observations, Phys. Rev. D **74**, 127305 (2006)
31.42 S. del Campo, R. Herrera: Curvaton field and intermediate inflationary universe model, Phys. Rev. D **76**, 103503 (2007)
31.43 S. del Campo: Approach to exact inflation in modified Friedmann equation, J. Cosmol. Astropart. Phys. **12**, 005 (2012)
31.44 S.M. Carroll, V. Duvvuri, M. Trodden, M. Turner: Is cosmic speed – Up due to new gravitational physics?, Phys. Rev. D **70**, 043528 (2004)
31.45 C. Gao: Generalized modified gravity with second order acceleration equation, Phys. Rev. D **86**, 103512 (2012)
31.46 R.-G. Cai, L.-M. Cao, Y.-P. Hu: Corrected entropy-area relation and modified friedmann equations, J. High Energy Phys. **08**, 090 (2008)
31.47 J.E. Lidsey: Thermodynamics of anomaly-driven cosmology, Class. Quantum Gravity **26**, 147001 (2009)
31.48 A.P. Apostolopoulos, G. Siopsis, N. Tetradis: Cosmology from an AdS Schwarzschild black hole via holography, Phys. Rev. Lett. **102**, 151301 (2009)
31.49 J.E. Lipsey, G. Siopsis, N. Tetradis: Holographic cosmology from the first law of thermodynamics and the generalized uncertainty principle, Phys. Rev. D **88**, 103519 (2013)
31.50 F. Gomez, P. Minning, P. Salgado: Standard cosmology in Chern-Simons gravity, Phys. Rev. D **84**, 063506 (2011)
31.51 T. Shiromizu, K. Maeda, M. Sasaki: The Einstein equation on the 3-brane world, Phys. Rev. D **62**, 024012 (2000)
31.52 A. Kamenshchik, U. Moschella, V. Pasquier: An Alternative to quintessence, Phys. Lett. B **511**, 265 (2001)
31.53 N. Bilic, G.B. Tupper, R.D. Viollier: Unification of dark matter and dark energy: The Inhomogeneous Chaplygin gas, Phys. Lett. B **535**, 17 (2002)
31.54 S. Dodelson, W.H. Kinney, E.W. Kolb: Cosmic microwave background measurements can discriminate among inflationmodels, Phys. Rev. D **56**, 3207 (1997)
31.55 A. Albrecht, P.J. Steinhardt: Cosmology for grand unified theories with radiatively induced symmetry breaking, Phys. Rev. Lett. **48**, 1220 (1982)
31.56 A.D. Linde: A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems, Phys. Lett. B **108**, 389 (1982)
31.57 A.D. Linde: Chaotic inflation, Phys. Lett. B **129**, 177 (1983)

31.58 D.H. Lyth: Particle physics models of inflation, Lecture Notes in Physics **738**, 81–118 (2008)
31.59 A.D. Linde: Hybrid inflation, Phys. Rev. D **49**, 748 (1994)
31.60 S. Mukohyama: Horava-Lifshitz cosmology: A review, Class. Quantum Gravity **27**, 223101 (2010)
31.61 A.R. Liddle, P. Parsons, J.D. Barrow: Formalizing the slow roll approximation in inflation, Phys. Rev. D **50**, 7222 (1994)
31.62 D.S. Salopek, J.R. Bond: Nonlinear evolution of long wavelength metric fluctuations in inflationary models, Phys. Rev. D **42**, 3936 (1990)
31.63 V.F. Mukhanov: Inflation: homogeneous limit, arXiv.org e-print archives: astro-ph/0511570 (2005)
31.64 J.D. Barrow, A.R. Liddle: Perturbation spectra from intermediate inflation, Phys. Rev. D **47**, 5219 (1993)
31.65 J.D. Barrow: Exact inflationary universes with potential minima, Phys. Rev. D **49**, 3055 (1994)
31.66 J. Magueijo, L. Smolin: Gravity's rainbow, Class. Quantum Gravity **21**, 1725 (2004)
31.67 L. Kai, Y. Shu-Zheng: An inflationary solution of scalar field in Finsler universe, Chin. Phys. Lett. **25**, 2382 (2008)
31.68 L. Kai, Y. Shu-Zheng: Exact scalar field inflationary solution in rainbow universe, Int. J. Theor. Phys. **47**, 2991 (2008)
31.69 L. Kai, Y. Shu-Zheng: A model with exact inflationary solution in Finsler universe, Int. J. Theor. Phys. **48**, 1882 (2009)
31.70 W.-F. Wang: Exact solution in chaotic inflation model with potential minima, Commun. Theor. Phys. (Beijing China) **36**, 122 (2001)
31.71 V.N. Lukash: Production of phonons in an isotropic universe, Pisma. Zhurnal Eksp. Teor. Fiz. **79**, 1601 (1980)
31.72 V.F. Mukhanov, G.V. Chibisov: Quantum fluctuation and nonsingular universe, J. Exp. Theor. Phys. Lett. **33**, 532 (1981)
31.73 S.W. Hawking: The development of irregularities in a single bubble inflationary universe, Phys. Lett. B **115**, 295 (1982)
31.74 A.A. Starobinsky: Dynamics of phase transition in the new inflationary universe scenario andgeneration of perturbations, Phys. Lett. B **117**, 175 (1982)
31.75 L.P. Grishchuk: Amplification of gravitational waves in an istropic universe, Sov. Phys. J. Exp. Theor. Phys. **40**, 409 (1975)
31.76 A.A. Starobinsky: Relict gravitation radiation spectrum and initial state of the universe, J. Exp. Theor. Phys. Lett. **30**, 682 (1979)
31.77 V. Rubakov, M. Sazhin, A. Veryaskin: Graviton creation in the inflationary universe and the grand unification scale, Phys. Lett. B **115**, 189 (1982)
31.78 R. Fabbri, M.D. Pollock: The effect of primordially produced gravitons upon the anisotropy of the cosmological microwave background radiation, Phys. Lett. B **125**, 445 (1983)
31.79 L. Abbott, M. Wise: Constraints on generalized inflationary cosmologies, Nucl. Phys. B **244**, 541 (1984)
31.80 V.F. Mukhanov, H. Feldman, R.H. Brandenberger: Theory of cosmological perturbations. Part 1. Classical perturbations. Part 2. Quantum theory of perturbations. Part 3. Extensions, Phys. Rep. **215**, 203 (1992)
31.81 V.N. Lukash: Formation of Large Scale Structure of the Universe. In: *VIII Brazilian School of Cosmology and Gravitation II*, ed. by M. Novello (Editions Frontiers, Rio de Janeiro 1995)
31.82 M. Kamionkowski, A. Kosowsky, A. Stebbins: A probe of primordial gravity waves and vorticity, Phys. Rev. Lett. **78**, 2058 (1997)
31.83 L. Knox, Y. Song: A Limit on the detectability of the energy scale of inflation, Phys. Rev. Lett. **89**, 011303 (2002)
31.84 Planck Collboration, Planck 2013 results. XXII. Constraints on inflation (2013), arXiv:1303.5082
31.85 M. Kamionkowski, A. Kosowsky: The Cosmic microwave background and particle physics, Annu. Rev. Nucl. Part. Sci. **49**, 77 (1999)
31.86 J.D. Barrow, P. Parsons: Inflationary models with logarithmic potentials, Phys. Rev. D **52**, 5576 (1995)
31.87 V.F. Mukanov: Gravitational instability of the universe filled with a scalar field, J. Exp. Theor. Phys. Lett. **41**, 493 (1985)
31.88 V.F. Mukanov: Gravitational instability of the universe filled with a scalar field, Pisma Zhurnal Eksp. Teor. Fiz. **41**, 402 (1985)
31.89 V.F. Mukanov: *Fisical Foundations of Cosmology* (Cambridge Univ. Press, Cambridge 2005)
31.90 M. Sasaki: Large scale quantum fluctuations in the inflationary universe, Prog. Theor. Phys. **76**, 1036 (1986)
31.91 J.M. Bardeen, P.J. Steinhardt, M.S. Turner: Spontaneous creation of almost scale – Free density perturbations in an inflationary universe, Phys. Rev. D **28**, 679 (1983)
31.92 E.D. Stewart, D.H. Lyth: A more accurate analytic calculation of the spectrum of cosmological perturbations produced during inflation, Phys. Lett. B **302**, 171 (1993)
31.93 D.H. Lyth, E.D. Stewart: The Curvature perturbation in power law (e.g. extended) inflation, Phys. Lett. B **274**, 168 (1992)
31.94 W.H. Kinney: A Hamilton–Jacobi approach to nonslow roll inflation, Phys. Rev. D **56**, 2002 (1997)
31.95 S. Kundu, J. Cosmol: Inflation with general initial conditions for scalar perturbations, Astropart. Phys. **1202**, 005 (2012)
31.96 A. Guth, S.-Y. Pi: Fluctuations in the new inflationary universe, Phys. Rev. Lett. **49**, 1110 (1982)

31.97 S. Mollerach, S. Matarrese, F. Lucchin: Blue perturbation spectra from inflation, Phys. Rev. D **50**, 4835 (1994)

31.98 J.-F. Dufaux, J.E. Lidsey, R. Maartens, M. Sami: Cosmological perturbations from brane inflation with a Gauss–Bonnet term, Phys. Rev. D **70**, 083525 (2004)

31.99 A.A. Starobinsky: Cosmic background anisotropy induced by isotropic flat-spectrum gravitational-wave perturbations, Sov. Astron. Lett. **11**, 133 (1985)

31.100 L. Hui, W.H. Kinney: Short distance physics and the consistency relation for scalar and tensor fluctuations in the inflationary universe, Phys. Rev. D **65**, 103507 (2002)

31.101 R. Easther, B.R. Greene, W.H. Kinney, G. Shiu: A Generic estimate of transPlanckian modifications to the primordial power spectrum in inflation, Phys. Rev. D **66**, 023518 (2002)

31.102 M.B. Hoffman, M.S. Turner: Kinematic constraints to the key inflationary observables, Phys. Rev. D **64**, 023506 (2001)

31.103 W.H. Kinney: Inflation: Flow, fixed points and observables to arbitrary order in slow roll, Phys. Rev. D **66**, 083508 (2002)

31.104 R. Easther, W.H. Kinney: Monte Carlo reconstruction of the inflationary potential, Phys. Rev. D **67**, 043511 (2003)

31.105 S. Chongchitnan, G. Efstathiou: Dynamics of the inflationary flow equations, Phys. Rev. D **72**, 083520 (2005)

31.106 A.R. Liddle: Inflationary flow equations, Phys. Rev. D **68**, 103504 (2003)

31.107 M. Spaliński: New solutions of the inflationary flow equations, J. Cosmol. Astropart. Phys. **0708**, 016 (2007)

31.108 P. Adshead, R. Easther: Constraining inflation, J. Cosmol. Astropart. Phys. **0810**, 047 (2008)

31.109 R. Easther: An Inflationary model with an exact perturbation spectrum, Class. Quantum Gravity **13**, 1775 (1996)

31.110 E. Gunzig, R. Maartens, A.V. Nesteruk: Inflationary cosmology and thermodynamics, Class. Quantum Gravity **15**, 923 (1998)

31.111 Z.-K. Guo, H.-S. Zhang, Y.-Z. Zhang: Inflationary attractor in the braneworld scenario, Phys. Rev. D **69**, 063502 (2004)

31.112 WMAP Collaboration: Seven-year Wilkinson microwave anisotropy probe (WMAP) observations: Cosmological interpretation, Astrophys. J. Suppl. **192**, 18 (2011)

31.113 M. Hicken, P. Challis, S. Jha, R.P. Kirshner, T. Matheson, M. Modjaz, A. Rest, W.M. Wood-Vasey, G. Bakos, E.J. Barton, P. Berlind, A. Bragg, C. Briceno, W.R. Brown, N. Caldwell, M. Calkins, R. Cho, L. Ciupik, M. Contreras, K.-C. Dendy, A. Dosaj, N. Durham, K. Eriksen, G. Esquerdo, M. Everett, E. Falco, J. Fernandez, A. Gaba, P. Garnavich, G. Graves, P. Green, T. Groner, C. Hergenrother, M.J. Holman, V. Hradecky, J. Huchra, B. Hutchison, D. Jerius, A. Jordan, R. Kilgard, M. Krauss, K. Luhman, L. Macri, D. Marrone, J. McDowell, D. McIntosh, B. McNamara, T. Megeath, B. Mochejska, D. Munoz, J. Muzerolle, O. Naranjo, G. Narayan, M. Pahre, W. Peters, D. Peterson, K. Rines, B. Ripman, A. Roussanova, R. Schild, A. Sicilia-Aguilar, J. Sokoloski, K. Smalley, A. Smith, T. Spahr, K.Z. Stanek, P. Barmby, S. Blondin, C.W. Stubbs, A. Szentgyorgyi, M.A.P. Torres, A. Vaz, A. Vikhlinin, Z. Wang, M. Westover, D. Woods, P. Zhao: CfA3: 185 type Ia supernova light curves from the CfA, Astrophys. J. **700**, 331 (2009)

31.114 WMAP Collaboration: Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Sky maps, systematic errors, and basic results, Astrophys. J. Suppl. **192**, 14 (2011)

31.115 X. Chen, M.X. Huang, S. Kachru, G. Shiu: Observational signatures and non-Gaussianities of general single field inflation, J. Cosmol. Astropart. Phys. **0701**, 002 (2007)

32. Cosmology with the Cosmic Microwave Background

Tarun Souradeep

The *standard* model of cosmology must not only explain the dynamics of the homogeneous background universe, but also satisfactorily describe the perturbed universe – the generation, evolution and finally, the formation of large-scale structures in the universe. Cosmic microwave background (CMB) has been by far the most influential cosmological observation driving advances in current cosmology. Exquisite measurements from CMB experiments have seen the emergence of a *concordant* cosmological model. Besides precise determination of various parameters of the *standard* cosmological model, observations have also established some important basic tenets that underlie models of cosmology and structure formation in the universe. The article reviews this aspect of recent progress in cosmology for a general science reader.

32.1 Contemporary View of our Cosmos

The universe is the grandest conceivable scale on which the human mind can strive to understand nature. The amazing aspect of cosmology, the branch of science that attempts to understand the origin and evolution of the universe, is that it is largely comprehensible by applying the same basic laws of physics that we use for other branches of physics. Historically, theoretical developments always preceded observations in cosmology up until the past couple of decades. Recent developments in cosmology have been largely driven by huge improvement in quality, quantity, and the scope of cosmological observations.

We will avoid giving a historical perspective. The theoretical model of cosmology, the Hot Big Bang model (HBBM), has broadly remained as it was established and widely accepted by the late 1960s. This is readily available in most standard textbooks, as well as, many semipopular books. The perspective would be to review the theoretical model of cosmology in the light of the available data. The main goal is to convey the excitement in cosmology where amazing observations have now concretely verified that the present edifice of the standard cosmological models is robust. A set of foundation and pillars of cosmology have emerged and are each supported by a number of distinct observations:

- Homogeneous, isotropic cosmology, expanding from a hot initial phase due to gravitational dynamics of the Friedmann equations derived from laws of general relativity.
- The basics constituent of the universe are baryons, photons, neutrinos, dark matter, and dark energy (cosmological constant/vacuum energy).
- The homogeneous spatial sections of spacetime are nearly geometrically flat (Euclidean space).
- Evolution of density perturbations under gravitational instability has produced the large-scale structure in the distribution of matter starting from the primordial perturbations in the early universe.
- The primordial perturbations have correlation on length scales larger than the causal horizon that makes a strong case for an epoch of inflation in the very early universe. The nature of primor-

dial perturbation matches that of the generation of primordial perturbations in the simplest model of inflation.

The cosmic microwave background (CMB), a nearly uniform, thermal black-body distribution of photons throughout space, at a temperature of 2.7 K, accounts for almost the entire radiation energy density in the universe. Tiny variation of temperature and linear polarization of these black-body photons of the cosmic microwave background arriving from different directions in the sky faithfully encodes information about the early universe and have traveled unimpeded across the observable universe, making them an excellent probe of the universe.

There are two distinct aspects to modern day cosmology – the background universe and the perturbed universe. The *standard* model of cosmology must not only explain the dynamics of the homogeneous background universe, but also satisfactorily describe the perturbed universe – the generation, evolution, and, finally, the formation of large-scale structures in the universe. It is fair to say that cosmology over the past few decade has increasingly been more dominated by the interplay between the theory and observations of the perturbed universe – the origin and evolution of large-scale structures in the matter distribution. The past few years have seen the emergence of a *concordant* cosmological model that is consistent both with observational constraints from the background evolution of the universe as well that from the formation of large-scale structures (LSS) in the universe. In particular, the much talked about dawn of *precision* era of cosmology has been ushered in by the study of the perturbed universe. Measurements of CMB anisotropy and polarization have been by far the most influential cosmological observation driving advances in current cosmology in this direction.

32.2 The Smooth Background Universe

In recent years, vast cosmological surveys have provided a three-dimensional map of the distribution of millions of galaxies extending to a billion light-years around us. If theorists were to start building a model of cosmology today, this would be the cosmos they would need to explain. As shown in Fig. 32.1, there is a rich organized structure in the distribution of galaxies in a region of about 100 Mpc. However, this is a typical (statistically speaking) sample of mass distribution. In other words, the mass distribution in the universe averaged over regions of a few hundred mega-parsecs is fairly uniform. A stronger case for the homogeneous cosmology actually comes from the high degree of uniformity in the temperature of the CMB. These provide observational support for the cosmological principle that postulates a homogeneous universe invoked by theorists in the 1920 to 1930s to develop the first physical models of cosmology.

The evolution of the universe is an initial value problem in general relativity that governs Einstein's theory of gravitation – the dynamical evolution in time of the three-dimensional spatial sections in the foliation of spacetime. The, now observationally confirmed, large-scale homogeneity and isotropy of the matter distribution implies that the spatial sections of the universe are homogeneous (i. e., 3-D spaces of constant curvature). This reduces the problem to one of the simplest applications of general relativity formulated as a dynamical system. The dynamics of the spatial sections reduces to the time evolution of the scale factor $a(t)$ of the spatial section. Averaged on large scales, the spatial sections at any time t are simply a scaled version of the present universe at time t_0 – i. e., the physical distance between two points in the universe at time t is given by $a(t)d$, where d is the present distance. It is convenient to define $a(t_0) = 1$ with no loss of generality. Observationally, the expansion of the universe causes a redshift $z = (1 - a)/a$, of the spectrum light from a (cosmologically) distant astrophysical object (galaxy or quasar) emitted at a time t, when the universe had a scale factor a. The observation that all galaxies (on the average) appear to have a redshift in the spectra proportional to their distance confirms the expansion of the universe. The cosmic time t, the scale factor $a(t)$, and the redshift z can be used interchangeably to label spatial hyper-surfaces of the evolving universe.

The dynamics of the universe is encoded in the simple Friedmann equation

$$\begin{aligned} H^2(t) &\equiv \left(\frac{\dot{a}}{a}\right)^2 \\ &= \frac{8\pi G}{3}\rho_c(\Omega_m + \Omega_r + \Omega_\Lambda + \Omega_K)\,, \end{aligned} \tag{32.1}$$



Fig. 32.1 The figure depicts the typical structure in the three-dimensional distribution of galaxies in the universe using a 100 Mpc sized cube carved out from the 2def Galaxy Redshift Survey (2dFGRS). The locations of galaxies are marked by a *toy* image colored according to the galaxy type. The gray shading is a visual aid that highlights the density contrast in the distribution by marking a region of approximately constant density (courtesy Paul Bourke/Swinburne Centre for Astrophysics and Supercomputing and the 2dFGRS Team)

deduced from the Einstein equations. It relates the Hubble parameter $H(t)$, that measures the expansion rate of the universe, to the matter density in the universe. Here we use the conventional dimensionless density parameter $\Omega_i = \rho_i/\rho_c$ in terms of the critical density $\rho_c = 3H^2/8\pi G$ at that time. The key components of the universe are radiation Ω_r, pressure-less gravitating matter Ω_m, and cosmological vacuum (dark) energy Ω_Λ. The departure of the total matter density parameter from unity contributes to the curvature of the space and can, hence, be represented by an effective curvature energy density Ω_K that determines the effect of curvature on the expansion of the universe. (Note that Ω_K is only a convenient notation and not a physical energy density, in particular, the *curvature density* is negative when the spatial section of uniform positive curvature.) Dividing out (32.1) by H^2 on both sides leads to a simple sum rule that summarizes the evolution of the universe

$$\Omega_m + \Omega_r + \Omega_\Lambda + \Omega_K = 1 . \qquad (32.2)$$

Since the expansion rate $H(t)$ evolves with time, Ω_i are time dependent. Further, the components (species) of matter are assumed to be noninteracting (on cosmological scales), ideal, hydrodynamic fluids, specified by their energy/mass density ρ_i and the pressure p_i (equivalently, by the equation of state w_i, where $p_i = w_i\rho_i$). For given species, the evolution of the density ρ_i is governed by the conservation equation of the energy–momentum tensor. In a volume V_0 in the current universe, the conservation equation implies

$$\begin{aligned} d(\rho_i a^3 V_0) + p_i(3a^2)V_0 = 0\,, \quad \text{or} \\ \mathrm{d}E + p\,\mathrm{d}V = 0 \end{aligned} \qquad (32.3)$$

where in arriving at the second equation we use the fact that the physical volume $V = V_0 a^3$. The second equation resembles the first law of thermodynamic for an isentropic system with energy E and work done under pressure p (recall, $\mathrm{d}E + p\,\mathrm{d}V = T\,\mathrm{d}S$). It is straightforward to derive the scaling of the energy density ρ_i with the evolution of the universe as relative to its present value ρ_{0i} as

$$\frac{\rho_i}{\rho_{0i}} = a^{-3(1+w_i)} . \qquad (32.4)$$

The equation of the state characterizes the ideal cosmological fluid, e.g., $w = 1/3$ for radiation (relativistic matter); $w = 0$ for pressure-less (nonrelativistic matter), curvature density can be expressed as an ideal fluid with $w = -1/3$, and $w = -1$ for a cosmological constant (in general, for the dark energy component $w < -1/3$).

The entire dynamics of the universe is then completely determined by the present matter composition. Explicitly, (32.1) and (32.4) lead to the more commonly seen version of the Friedmann equation

$$\begin{aligned} H^2(t) &\equiv \left(\frac{\dot{a}}{a}\right)^2 \\ &= \frac{8\pi G}{3}\rho_{0c}\left[\Omega_{0m}a^{-3} + \Omega_{0r}a^{-4}\right. \\ &\quad \left. + \Omega_\Lambda + \Omega_{0K}a^{-2}\right] . \end{aligned} \qquad (32.5)$$

Equation (32.5) shows that the energy in an expanding universe is dominated successively by matter with a smaller value of w – i. e., first a radiation dominated phase Ω_r, followed by *matter-dominated* Ω_m, then curvature-dominated Ω_K and finally a cosmological vacuum (dark) energy Ω_Λ.

The relativistic mater density is almost entirely dominated by the CMB and the relic background of three species of light neutrinos (expected to have a density 68% of that of the CMB). The isentropic expansion dictated by the Friedmann equations implies that although, at present (given by the temperature of the CMB), Ω_{0r} is negligible, at an early epoch the universe was dominated by relativistic matter density. The

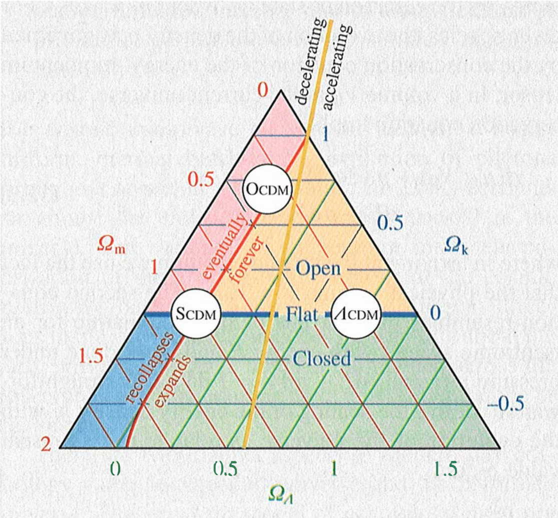


Fig. 32.2 The *cosmic triangle* represents the three key cosmological parameters – Ω_m, Ω_Λ, and Ω_k – where each point in the triangle satisfies the sum rule $\Omega_m + \Omega_\Lambda + \Omega_k = 1$. The *blue horizontal line* (*marked Flat*) corresponds to a flat universe ($\Omega_m + \Omega_\Lambda = 1$), separating an open universe from a closed one. The red line, nearly along the $\Lambda = 0$ line, separates a universe that will expand forever (approximately $\Omega_\Lambda > 0$) from one that will eventually recollapse (approximately $\Omega_\Lambda < 0$). And the *yellow, nearly vertical line* separates a universe with an expansion rate that is currently decelerating from one that is accelerating. The locations of three key models are highlighted: (Flat) standard cold dark matter (SCDM); flat (ΛCDM – Lambda-cold dark matter); and open cold dark matter (OCDM) (after [32.1, 2])

pressure-less matter density $\Omega_m = \Omega_B + \Omega_{cdm} + \Omega_\nu$ minimally consists of three distinct components, the baryonic matter, cold dark matter, and a possibly minor contribution from massive neutrino species. The constraint on the Baryon density $\Omega_B h^2 = 0.022 \pm 0.002$ from the predicted abundances of light elements from Big-Bang nucleosynthesis (BBN) is consistent with that recently obtained from considerations of structure formation.

The present state of the universe in terms the three dominant components can be neatly summarized on the *cosmic triangle* shown in Fig. 32.2 [32.2]. The three axes address fundamental issues regarding background cosmology – Does space have positive, negative or zero curvature (Ω_{0K})? Is the expansion accelerating, or decelerating (determined by Ω_Λ)?, and, what is the fraction of the nonrelativistic matter, (Ω_{0m})?

Historically, the focus has shifted between different sectors of the cosmic triangle depending on which of the three is the dominant player Ω_{0m}, Ω_{0K}, or, Ω_Λ. The canonical standard cold dark matter (SCDM) is a model where the present universe is a flat universe ($\Omega_{0K} = 0$) dominated by nonrelativistic matter density $\Omega_{0m} = 1$ ($\Rightarrow \Omega_\Lambda = 0$). This is also theoretically the simplest since it avoids the fine tuning problem of having a curved universe by invoking inflation and was the favorite in the 1980s. The nonrelativistic matter had to be mostly nonbaryonic dark matter (i. e., matter than does not interact with light), since Big-Bang nucleosynthesis and the absence of CMB temperature fluctuations at the power level of $\approx 10^{-4}$ limit the baryonic fraction to a much smaller value than that inferred for Ω_{0m}. (Nonbaryonic dark matter component has to be nonrelativistic to satisfy power spectrum measurements of the LSS.)

At the end of the 1980s and early 1990s, observations of LSS made it clear that Ω_{0m} was much smaller than unity. The sum rule, (32.2), then implies that either Ω_{0K}, or Ω_Λ, or both had to be non zero. The theoretical discomfort with a nonzero Ω_Λ (that still persists today) led to the era of open cold dark matter (OCDM) models, where $\Omega_{0K} > 0$. The conflict $\Omega_{0K} \neq 0$ with a robust prediction of inflation promptly development of *open* models of inflationary scenarios that could avoid this problem.

Toward the end of the 1990s, the observation of a high-redshift supernova indicated an acceleration in the present expansion universe. Very soon after, CMB anisotropy observations revealed a flat universe ($\Omega_{0K} = 0$). This leads to the currently favored ΛCDM model in cosmology. The energy density of the cosmological constant (or, more broadly quintessence) can be inferred from the measurement of luminosity distance as a function of redshift using the high-redshift supernova SN Ia as standard candles. In this chapter, we limit our attention to the simplest case of a cosmological constant that has a constant equation of state $w = -1$, which is also completely consistent with all observations to date. Alternative propositions for the nature of the dark energy are discussed in the chapter by Tsujikawa in this volume.

The key program of the Hubble space telescope (HST) mission measured the expansion rate of the universe $H_0 = 72 \pm 8\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ in 2001. Recently, new Spitzer calibration, has allowed the systematic uncertainty in H_0 from the HST key project to be decreased by over a factor of 3. Also optical and infrared observations of over 600 Cepheid variables in

the host galaxies of eight recent Type Ia supernovae (SNe Ia) determines $H_0 = 73.8 \pm 2.4\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. This is broadly consistent with the constraints from the CMB anisotropy and large-scale structure observations and combined constraints are remarkably tight. Cosmological observations have definitively determined the present universe to be located in the Λ-CDM sector. (The above improvement in H_0, combined with Wilkinson microwave anisotropy probe (WMAP)-7yr data, results in a strong constraint on the nature of dark energy $w = -1.08 \pm 0.10$, close to a cosmological constant.) The expansion rate and age estimates of the present universe measured from CMB data are again consistent with, and considerably improved in precision by including structure formation consideration.

One the most crucial observational pillars that support the HBBM of the background universe is the CMB discussed in the next section.

32.3 The Cosmic Microwave Background

The CMB, a nearly uniform, thermal black-body (Planck) distribution of photons throughout space, at a temperature of 2.7 K, accounts for almost the entire radiation energy density in the universe. The HBBM ascribes cosmic significance to this microwave radiation background, and hence CMB measurements play a role of great importance. In this widely accepted view, the CMB comprises the oldest photons that last interacted when the universe was only 380 000 yr old (compared to the present age of about 14 billion years). The photons have freely traveled right from the edge of the observable universe a distance of about 43 billion light years (14 Gpc) as explained in Fig. 32.3.

The prediction of the Planck distribution of the CMB in the HBBM dates from the early nucleosynthesis calculations of Gamow and collaborators in 1948. Thermal equilibrium in the early universe establishes a Planck energy distribution for the photons. In the HBBM the universe expands adiabatically conserving the photon entropy per comoving volume. (The observed CMB accounts for almost all the entropy.) The adiabatic Hubble expansion conserves the Planck distribution. However, the energy density of photons $\rho_r \propto a^{-4}$ in an expanding universe (see (32.4) for radiation $w = 1/3$). Recalling, that the energy density of a black body is proportional to the fourth power of the temperature, it is clear that the temperature of the CMB photons $T_{\mathrm{cmb}} = T_{0\mathrm{cmb}}/a = T_{0\mathrm{cmb}}(1+z)$ scales as the inverse of the expansion of universe. At a redshift of $z_{\mathrm{rec}} \approx 1100$, the temperature of CMB falls below the threshold required to keep the hydrogen atoms in the universe ionized. At this epoch of *recombination* at around $t = 380\,000$ yr, the protons and electrons form a neutral hydrogen atom and lose their coupling to the CMB. (This happens a bit earlier for the helium fraction). The baryonic matter in the universe transits from an ionized plasma state to neutral one where CMB photons can freely travel over cosmic distances.

The serendipitous discovery of this extra galactic microwave background Penzias and Wilson in 1965 provided a big boost to the HBBM. This was fol-

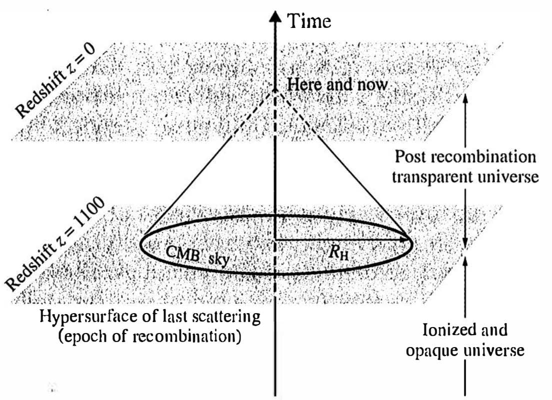


Fig. 32.3 A cartoon explaining the CMB using a space-(conformal) time diagram. The universe became transparent at the epoch of recombination, and CMB photons were able to travel to us freely over cosmic distances along our past light cone. In an expanding universe, the temperature of the Planck black-body CMB is inversely proportional to the expansion factor. When the universe is about 1100 times smaller, the CMB photons are hot enough to keep the baryonic matter in the universe (about three quarters hydrogen, one quarter helium as determined by Big-Bang nucleosynthesis) ionized, accompanied by a sharp transition to an opaque universe. The CMB photons come to us unimpeded directly from this spherical opaque surface of last scattering at a distance of $R_H = 14$ Gpc that surrounds us – a super IMAX cosmic screen. The *brown circle* depicts the sphere of last scattering in the reduced 2 + 1-dimensional representation of the universe

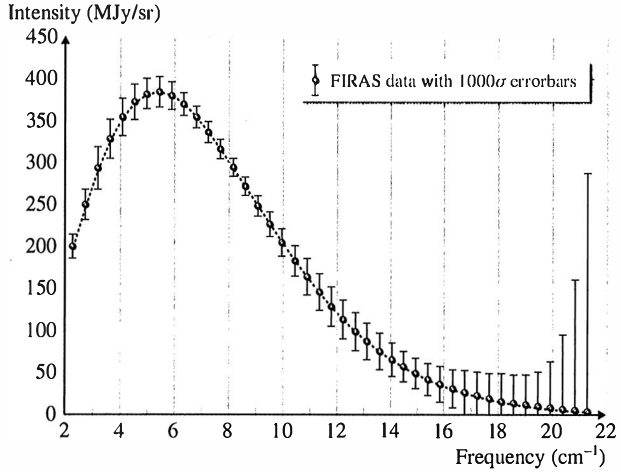


Fig. 32.4 Measurements of the energy spectrum of the CMB photons as a function of frequency (from 60 to 600 GHz). The measurements are from the FIRAS instrument on board the COBE satellite that won the Nobel prize in Physics in 2006. The accuracy of the measurements is apparent from the fact that the error bars have been multiplied by a factor of thousand in the plot. The distribution is extremely well fit by a black-body spectrum at a temperature of $T_0 = 2.726$ (± 0.0013) making the CMB the most perfect black body known in nature (courtesy of Tuhin Ghosh (IUCAA))

lowed up by numerous measurements of the CMB flux at other wavelengths that were broadly consistent with a Planck distribution of CMB photons. The Nobel prize in Physics in 2006 was awarded to John Mather (NASA Goddard Flight Center, USA) and George Smoot (University of Berkeley, USA), who led experimental teams of the pioneering Cosmic Background Explorer (COBE) mission – a US space Administration, NASA, satellite launched in 1989 to measure the cosmic microwave background radiation with unprecedented accuracy over the full sky. The satellite operated for 4 yr in a circumpolar orbit at an altitude of 900 km. COBE carried three different instruments: far-infrared absolute spectrophotometer (FIRAS), differential microwave radiometer (DMR), and diffuse infrared background experiment (DIRBE). John Mather was the principle investigator (PI of the FIRAS experiment that measured the energy distribution of CMB photons to unprecedented accuracy. The FIRAS instrument measurements of the radiation flux in the 60–2880 GHz frequency band shown in Fig. 32.4 confirmed the Planck distribution of CMB photons beyond reasonable doubt. The flux measurement at a given wavelength can be converted into an equivalent thermodynamic temperature T_0 for the CMB. Recent results derived from the FIRAS data combined with WMAP in 2009 find that the energy spectrum of CMB photons is accurately described by a Planck distribution at the precise temperature

$$T_0 = 2.726 \pm 0.0013\,\mathrm{K}\,. \tag{32.6}$$

Over the frequency band 60–630 GHz used to deduce the above FIRAS result, the maximum 1-σ deviation of the CMB spectrum from a Planck is constrained to be $\lesssim 0.01\%$ of the peak brightness. The observationally established Planck distribution of the energy spectrum of the CMB is naturally explained as arising from the thermal equilibrium the baryons and photons set up at very high temperatures and densities, that is expected to exist in the early universe. The present temperature T_0, of the CMB sets the total entropy of the universe (given the number of relativistic neutrino species). The origin of this entropy is not explained within the classical Big Bang model (inflation scenarios do provide an explanation but not a prediction). Working backward in time, adiabatic expansion implies a smaller and hotter universe expected in the HBBM.

32.4 Perturbed Universe: Structure Formation

The *standard* model of cosmology must not only explain the dynamics of the homogeneous background universe, but also describe the perturbed universe – the generation, evolution, and the formation of large-scale structures in the universe. There is a well understood (if not rigorously defined) notion of a *standard* model of cosmology that includes the formation of a large-scale structure. It is fair to say that much of the recent progress in cosmology has come from the interplay between refinements of the theories of structure formation and the improvements in observations.

Although the simple homogeneous and isotropic cosmological model does fit the dynamics of the background universe averaged on large scales, the rich structure in the distribution of galaxies shown in Fig. 32.1 suggests that there is more information to be gleaned about the universe from the large-scale structure of mass distribution (LSS). It has been a well-accepted

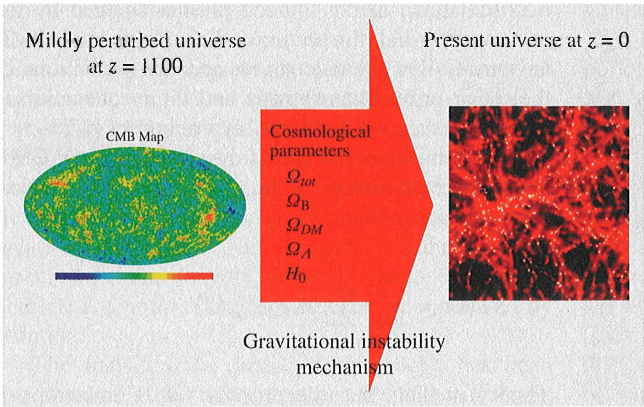


Fig. 32.5 A schematic figure to illustrate how understanding of the perturbed universe *determines* the cosmological parameters. The exquisitely measured CMB anisotropy maps characterize the mildly perturbed universe at early times. Large galaxy surveys and other LSS probe give the final state of the LSS. The cosmological parameters that affect the known structure formation mechanism through gravitational instability have to be dialed to precise values to consistently produce the LSS in the present universe from the mildly perturbed universe observed in the CMB anisotropy

notion that the large-scale structure in the distribution of matter in the present universe arose gradually due to gravitational instability from tiny primordial perturbation in the early universe. Although explosive mechanisms for structure formation in a relatively recent epoch had been proposed, the limits of input into the radiation budget in the recent past due to the tight adherence of the CMB to the Planck form seen in the COBE-FIRAS data make them nonviable. Also, the tiny level of fluctuations in the temperature of the CMB implies that the level of inhomogeneity in the universe at a redshift of $z_{\text{rec}} = 1100$ is at most few 10 ppm. A recent exciting success of observational cosmology has been in detecting the baryon acoustic oscillations that establish the gravitational instability mechanism beyond reasonable doubt.

As schematically summarized in Fig. 32.5, cosmological observations have placed the theory of structure formation in an enviable position for any branch of physics where the initial and final states as well as the dynamical mechanism are known:

- The exquisite maps of CMB anisotropy provide a snap shot of perturbation in the universe at a redshift of $z_{\text{rec}} = 1100$ when the universe is only about 380 000 yr old.
- In the past decade an extensive survey of galaxies has mapped out the distribution of matter in the present 14 Gyr-old universe.
- As mentioned above, the well-understood gravitationally instability is the underlying mechanism for amplifying the tiny perturbations at a redshift of $z_{\text{rec}} = 1100$ to give rise to the observed LSS now.

The recent era of *precision* cosmology arises from the sensitivity of a consistent picture on the cosmological parameters. The parameters have to be dialed to precise values to make a consistent description of the perturbed universe starting with the mildly perturbed universe at $z_{\text{rec}} = 1100$ seen in the CMB to the present universe with a well-developed LSS.

32.5 CMB Anisotropy and Polarization

The CMB photons arriving from different directions in the sky show tiny variations in temperature, at a level of ten parts per million, i. e., tens of micro-Kelvin, referred to as the CMB anisotropy, and a net linear polarization pattern at micro-Kelvin to tens of nano-Kelvin level. The tiny variation of temperature and linear polarization of these black-body photons of the cosmic microwave background arriving from different directions in the sky faithfully encodes information about the early universe and have traveled unimpeded across the observable universe making them an excellent probe of the universe. As illustrated in the cartoon in Fig. 32.3, the cosmic microwave background radiation sky is essentially a *giant, cosmic super IMAX theater screen* surrounding us at a distance of 43 billion light-years displaying a snapshot of the universe at a time very close to its origin.

The CMB anisotropy is imprints of the perturbed universe in the radiation when the universe was only 380 000 yr old. On the large angular scales, the CMB

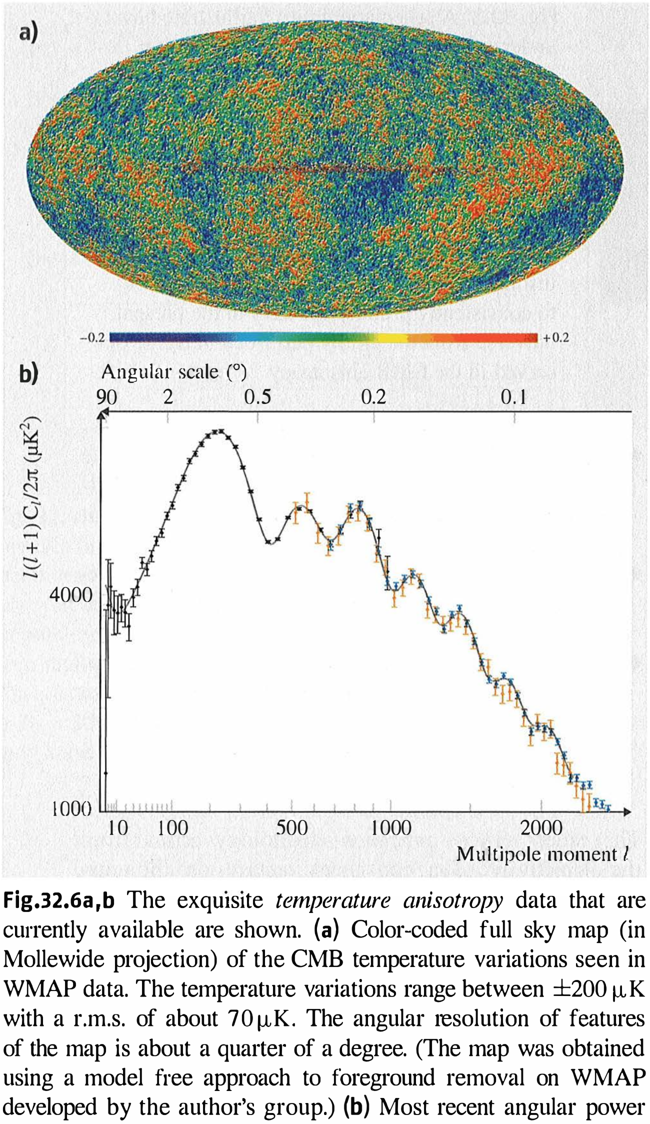


Fig.32.6a,b The exquisite *temperature anisotropy* data that are currently available are shown. **(a)** Color-coded full sky map (in Mollewide projection) of the CMB temperature variations seen in WMAP data. The temperature variations range between $\pm 200\,\mu$K with a r.m.s. of about $70\,\mu$K. The angular resolution of features of the map is about a quarter of a degree. (The map was obtained using a model free approach to foreground removal on WMAP developed by the author's group.) **(b)** Most recent angular power spectrum of CMB obtained from the entire WMAP 9 yr (*black*), the ground-based South Pole Telescope (*blue*), and Atacama Cosmology Telescope (*orange*) data. The solid gray curve shows that the best-fit power law, flat, ΛCDM model obtained from WMAP-9 threads all the data points closely (WMAP-9 publication publicly available at the NASA-GSFC LAMBDA site http://www.lambda.gsfc.nasa.gov)

anisotropy directly probes the primordial power spectrum on scales enormously larger than the *causal horizon*. On smaller angular scales, the CMB temperature fluctuations probe the physics of the coupled baryon–photon fluid through the imprint of the acoustic oscillations in the ionized plasma sourced by the same primordial fluctuations. The physics of CMB anisotropy is well understood, and the predictions of the linear primary anisotropy and their connection to observables are, by and large, unambiguous [32.3–5].

It is convenient to express the sky map of the CMB temperature anisotropy in the direction $\hat{\boldsymbol{n}}$ as a spherical harmonic expansion

$$\Delta T(\hat{\boldsymbol{n}}) = \sum_{\ell=2}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\hat{\boldsymbol{n}}) \,.$$

Theory predicts that the primary CMB anisotropy is a Gaussian field (of zero mean), and current observations remain fully consistent with this expectation. The anisotropy can then be characterized solely in terms an angular spectrum

$$C_\ell = \frac{1}{(2\ell+1)} \sum_{m=-\ell}^{\ell} |a_{\ell m}|^2 \,.$$

The C_ℓ spectra for a wide variety of models share a generic set of features clearly related to basics physics of primary CMB anisotropy.

The acoustic peaks occur because the cosmological perturbations excite acoustic waves in the relativistic plasma of the early universe. The recombination of baryons at redshift $z \approx 1100$ effectively decouples the baryon and photons in the plasma abruptly switching off the wave propagation. In the time between the excitation of the perturbations and the epoch of recombination, a sound wave could have traveled a fixed distance. Modes of different wavelengths can complete different numbers of oscillation periods. This translates the characteristic time into a characteristic length scale and produces a harmonic series of maxima and minima in the CMB anisotropy power spectrum. The acoustic oscillations have a characteristic scale known as the sound horizon, which is the comoving distance that a sound wave could have traveled up to the epoch of recombination. This is a well-determined physical scale imprinted on the CMB fluctuations on the surface of last scattering – the *cosmic super-IMAX* screen.

The angle subtended by this physical scale in the CMB anisotropy sky at the distance of 14 Gpc allows a sensitive determination of the geometry ($\Omega_{0\mathrm{K}}$) of the background universe. This is determined by the location of the harmonic peaks series of C_ℓ seen in Fig. 32.6. The amplitude of baryon–photon oscillations

consequently, the height of the peaks in the C_ℓ sensitively determine the baryon density Ω_B. The C_ℓ are sensitive to other important cosmological parameters, such as, the relative density of matter Ω_m; cosmological constant Ω_Λ; Hubble constant H_0, and deviation from flatness (curvature) Ω_K. Implicit in C_ℓ is the hypothesized nature of random primordial/initial metric perturbations – (Gaussian) statistics , (nearly scale invariant) power spectrum, (largely) adiabatic versus isocurvature, and (largely) scalar versus tensor component. The *default* settings in bracket are motivated by inflation.

The transition to precision cosmology has been spearheaded by the measurements of CMB anisotropy and, more recently, polarization. The COBE-DMR detection of CMB anisotropy provided observational evidence for the origin and mechanism of structure formation in the universe. The following decade has been dominated by high-resolution, full sky, CMB anisotropy measurements from the WMAP of NASA that has provided observational support for the basic acoustic physics of the baryon–photon plasma.

The measured angular power spectrum of the CMB temperature fluctuations C_ℓ, shown in Fig. 32.6 has become invaluable observables for constraining cosmological models. The position and amplitude of the peaks and dips of the C_ℓ are sensitive to important cosmological parameters. The most robust constraint obtained is that on the spatial curvature of the universe and baryon density. Combining most recent CMB observations from WMAP9, South Pole Telescope (SPT) and Atacama Cosmology Telescope (ACT) can establish that the space on cosmic scales is geometrically flat ($\Omega_K = 0.001 \pm 0.012$) to nearly within 1% precision. From WMAP9 alone, the dominant energy content in the present universe is a mysterious matter with negative pressure dubbed, dark energy, or a cosmological constant of about 72% ($\Omega_\Lambda = 0.721 \pm 0.025$), followed by cold nonbaryonic dark matter about 23% ($\Omega_m = 0.233 \pm 0.023$) and ordinary matter (baryons) account for only about 5% ($\Omega_B = 0.0463 \pm 0.00234$) of the matter budget. Observations of the large-scale structure in the distribution of galaxies, high-redshift supernova, and more recently, CMB polarization, have provided valuable complementary information.

In addition to the temperature anisotropy, there is also linear polarization information imprinted on the CMB at the last scattering surface. Thomson scattering generates CMB polarization anisotropy at decoupling. The coordinate-free description distinguishes two kinds of polarization patterns on the sky by their different parities. In the spinor approach, the even parity pattern is called the E-mode and the odd parity pattern the B-mode. While the CMB temperature anisotropy can also be generated during the propagation of the radiation from the last scattering surface, the CMB polarization signal can be generated only at the last scattering surface, where the optical depth transits from large to small values. The polarization information complements the CMB temperature anisotropy by isolating the effect at the last scattering surface from effects along the line of sight. Since the CMB polarization is sourced by the anisotropy of the CMB at the surface of last scattering, the angular power spectra of temperature and polarization are strongly linked to each other. For adiabatic initial perturbations, the acoustic peaks in the polarization spectra are out of phase with that of the temperature.

The Degree Angular Scale Interferometer (DASI) first measured the CMB polarization spectrum over a limited band of angular scales (multipole band $l \approx 200-440$) in late 2002. Since then, the polarization power spectrum measurements have been further refined by a number of CMB experiments, notably, MAXIMA CBI, QUaD, BICEP (background imaging of cosmic extragalactic polarization), etc. The main results indicated by the E-mode polarization measurements is that the acoustic peaks in the polarization spectra are indeed out of phase with that of the temperature. The strong limit on the nonadiabatic contribution to the primordial perturbations constrains the physics of the early universe.

The CMB polarization is a very clean and direct probe of the energy scale of early universe physics that generate the primordial metric perturbations. In the standard model, inflation generates both (scalar) density perturbations and (tensor) gravity wave perturbations. The relative amplitude of inflationary GW to scalar density perturbations sets the energy scale for inflation. A measurement of B-mode polarization on large-angular scales would give this amplitude, and hence a direct determination of the energy scale of inflation. Besides being a generic prediction of inflation, the cosmological gravity wave background from inflation would be a fundamental test of GR on cosmic scales and the semiclassical behavior of gravity.

32.6 Conclusion

The remarkable transition to precision cosmology has been spearheaded by the nearly two decade long experimental successes of CMB measurements. The first results from the COBE team (awarded the Nobel prize in Physics in 2006) provided only a coarse image of infant universe. The data from the Wilkinson Microwave Anisotropy Probe (WMAP) refined the image of the infant universe considerably in the following decade. It is the precision of these measurements of the CMB fluctuations cosmology that has translated to present day precision cosmology.

The past decade has seen the emergence of a *concordant* cosmological model that is consistent, both, with observational constraints from the background evolution of the universe, and that from the formation of a large-scale structure in the distribution of matter in the universe. Besides precise determination of various parameters of the *standard* cosmological model, CMB and related observations have also established some important basic tenets of cosmology and structure formation in the universe – *acausally* correlated initial perturbations, adiabatic nature primordial density perturbations, gravitational instability as the mechanism for structure formation. We have inferred a spatially flat universe where structures form by the gravitational evolution of nearly scale invariant, adiabatic perturbations, as expected from inflation.

The signature of primordial perturbations observed as the CMB anisotropy and polarization is the most compelling evidence for new, possibly fundamental, physics in the early universe that underlie the scenario of inflation (or related alternatives). Some fundamental *assumptions* rooted in the paradigm of inflation are still to be observationally established beyond doubt. Besides, there are deeper issues and exotic possibilities that no longer remain theoretical speculations, but have now come well within the grasp of cosmological observations (Chap. 39). These include cosmic topology, extra-dimensions, and violations of basic symmetries such as Lorentz transformations. In order to detect the subtle signatures it is also important to identify and weed out systematic effects such as the noncircularity of the beam in the acquisition and analysis of the CMB data.

The progress in the field continues unabated, refining the cosmological parameters into increasingly more precise numbers. Numerous ongoing and near future ground and balloon born CMB experiments at high sensitivity and resolution have sustained a steady pace of progress. The Planck Surveyor mission of ESA (European Space Agency) launched in May 2009 has acquired considerably more refined CMB measurements compared to WMAP. In the near future, exquisite results from the Planck satellite are expected. Planck is arguably the most ambitious cosmological space mission till date. It aims to measure CMB fluctuations at higher sensitivity and angular resolution to eke out almost all the information expected to be available in the CMB sky. Further in the future, dedicated CMB polarization space missions are being studied by both NASA and ESA [32.6].

References

32.1 J.P. Ostriker, T. Souradeep: The current status of observational cosmology, Pramana **63**, 817 (2004)

32.2 N. Bahcall, J.P. Ostriker, S. Perlmutter, P.J. Steinhardt: The cosmic triangle: Revealing the state of the universe, Science **284**, 1481 (1999)

32.3 W. Hu, S. Dodelson: Cosmic microwave background anisotropies, Annu. Rev. Astron. Astrophys. **40**, 171 (2002)

32.4 Pedagogical online material on CMB anisotropy and polarization available online at http://background.uchicago.edu/

32.5 Pedagogical online material on basic cosmology available online at http://www.astro.ucla.edu/~wright/cosmolog.htm

32.6 NASA/DOE/NSF Task Force: Report on Cosmic Microwave Background Research (2005), available online at http://www.nsf.gov/mps/ast/tfcr.jsp; also available at the Legacy Archive for Microwave Background Data analysis (LAMBDA) site http://lambda.gsfc.nasa.gov/

Part F Spacetim

Part F Spacetime Beyond Einstein

33. Quantum Gravity

Claus Kiefer

This chapter presents the motivations to quantize gravity and gives a brief introduction into the main approches.

33.1 Why Quantum Gravity?

33.1.1 Introduction

At the fundamental level of physical theories, we currently distinguish between four different interactions: strong, weak, electromagnetic, and gravitational interaction. In the standard model of particle physics, weak and electromagnetic interactions are partially unified in electroweak interaction, but otherwise they have so far been treated distinctly. The main difference lies between gravity and the other interactions. Whereas strong and electroweak interactions are successfully described by quantum field theories, all known gravitational phenomena can be understood by a classical theory: Einstein's theory of general relativity (GR). They range from applications in everyday life (such as the positioning system GPS) and the Solar System to stars, galaxies, and the Universe as a whole. The question is whether this is the final state of affairs or whether gravity must also be accommodated at the most fundamental level into the framework of quantum theory.

In this chapter, we first discuss the main arguments that can be invoked in favor of a quantum theory of gravity. We then present the main approaches and briefly discuss some of their applications. In order not to overload this article with too many references, we refer mainly to monographs and reviews, in which all references to the original articles can be found; this holds, in particular, for the monograph [33.1]. Some references to recently published original articles are, however, included.

Let us first look at Einstein's theory of GR. Gravity is there described by the geometry of space and time, unified to a four-dimensional manifold of spacetime. The theory can be defined by the Einstein–Hilbert action,

$$S_{\mathrm{EH}} = \frac{c^4}{16\pi G}\int_{\mathcal{M}} \mathrm{d}^4x\sqrt{-g}\,(R-2\Lambda) - \frac{c^4}{8\pi G}\int_{\partial\mathcal{M}} \mathrm{d}^3x\sqrt{h}K\,. \tag{33.1}$$

The first term is an integral over a spacetime region $\mathcal{M}$, while the second term is an integral over its spacelike boundary $\partial\mathcal{M}$; in it, K is the trace of the second fundamental form, and h is the determinant of the three-dimensional metric at the boundary. The need for this second term to obtain a consistent variational principle was already emphasized by Einstein in 1916.

In the presence of nongravitational fields, (33.1) is augmented by a *matter action* S_{m}. From the sum of these actions, one finds Einstein's field equations by variation,

$$G_{\mu\nu} := R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = \frac{8\pi G}{c^4}T_{\mu\nu} - \Lambda g_{\mu\nu}\,. \tag{33.2}$$

The right-hand side contains the symmetric (Belinfante) energy–momentum tensor

$$T_{\mu\nu} = \frac{2}{\sqrt{-g}} \frac{\delta S_{\mathrm{m}}}{\delta g^{\mu\nu}} \,, \tag{33.3}$$

plus the cosmological constant term, which may itself be accommodated into the energy–momentum tensor of the *vacuum*. If fermionic fields are added, one must generalize GR to the Einstein–Cartan theory or to the Poincaré gauge theory, because spin is the source of torsion, a geometric quantity that is identically zero in GR.

In his 1936 article on physics and reality, Einstein compared the left-hand side of (33.2) with fine marble and the right-hand side with rough timber. The reason is that he regarded the geometric part as already perfect, but the matter part with its many (not unified) contributions as a preliminary description. The attempts by Kaluza and later scientists to describe everything by going to a higher-dimensional spacetime can be seen as an attempt to overcome this dichotomy, which still persists today.

It is interesting to note that in particular the *timber part* is responsible for the incompleteness of GR. Natural conditions for the energy–momentum tensor lead to the *singularity theorems* according to which, for example, time-like geodesics reach their end in a finite proper time [33.2]. Important examples are the interior of black holes and the beginning (and possibly end) of the Universe.

The singularity theorems provide the first motivation for going beyond GR and constructing an encompassing, more fundamental, theory. We cannot understand the origin of the Universe and the final fate of black holes within GR. It is, of course, not logically required that the more fundamental theory must be a *quantum* theory. It is only by historic analogy (the avoidance of atomic instability in quantum mechanics) that this hope has arisen.

The second motivation for quantum gravity is a philosophical (and partly, historical) one. The reductionist program has proven to be very successful: hitherto unconnected theories such as optics, magnetism, and electricity have turned out to be different aspects of one and the same theory, electrodynamics. The standard model with its partial unification provides another example. Since gravity couples universally to all forms of energy, it couples to all quantum fields, which is why one expects that in a unified description of Nature the gravitational field is of quantum nature, too. Some scientists have expressed the hope that a quantum theory of gravity will not only cure the singularities of classical GR, but also the notorious divergences of quantum field theory. This may not come as a surprise, since both type of singularities are connected with the microstructure of spacetime, in particular to the open question whether it is fundamentally discrete or continuous.

Two further motivations are both of a conceptual nature. One is often called the *problem of time*. The point is that time is of very different nature in quantum (field) theory and in GR. In quantum mechanics, the t in the Schrödinger equation is Newton's absolute time; in quantum field theory, the fields act on Minkowski spacetime, which plays the role of a nondynamical background. In GR, there are no absolute structures; spacetime is fully dynamical. This is also called background independence. There is thus a clash of concepts. As long as the two frameworks are applied to different phenomena at different scales, this does not matter too much. If, however, one desires to understand phenomena where the interaction of gravity with quantum fields becomes dynamically relevant, one needs a conceptually coherent framework. One thus expects that a theory of quantum gravity will entail far-reaching consequences for the concept of time.

The other motivation was expressed, in particular, by Feynman during the famous Chapel Hill Conference in 1957. A quantum superposition, such as the superposition of two spin states for an electron in a Stern–Gerlach experiment will also lead to a superposition of gravitational fields; the components of the superposition correspond to different gravitational fields, which (at least in a gedanken experiment) can be transferred to different gravitational fields of a macroscopic object. As long as the superposition principle is valid also in these circumstances (which is not obvious but is a conservative assumption), one must invoke a quantum theory of gravity to describe this situation of superposed states of the gravitational field.

A final motivation comes from particle physics [33.3]. It seems that the standard model does not exist as a consistent quantum field theory up to arbitrarily high energies. The reasons for this failure may be a potential instability of the effective potential and/or potential Landau poles, although with the recently measured Higgs mass of about 126 GeV it is conceivable that the standard model holds up to the Planck scale (see below), where effects of quantum gravity are expected to come into play anyway. The standard model thus points to a more fundamental theory, which could be quantum gravity.

A main problem in the search for a quantum theory of gravity is the lack of empirical tests so far. Conceptually, effects of such a theory can occur at any scale, as long as the validity of the superposition principle is not restricted. The scale where one would definitely expect to see effects of quantum gravity is, however, far remote from the scales that are directly accessible by today's technology. It is the Planck scale. It was noted by Planck already in 1899 (1 year before his official introduction of the quantum of action!) that the fundamental constants of the speed of light (c), gravitational constant (G), and quantum of action (today called $\hbar$) can be combined in a unique way (apart from numerical factors) to give units of length, time, and mass. They are called the Planck length, l_P, Planck time, t_P, and Planck mass, m_P, respectively, and are given by the expressions

$$l_P := \sqrt{\frac{\hbar G}{c^3}} \approx 1.62 \times 10^{-33}\,\text{cm}\,, \tag{33.4}$$

$$t_P := \frac{l_P}{c} = \sqrt{\frac{\hbar G}{c^5}} \approx 5.39 \times 10^{-44}\,\text{s}\,, \tag{33.5}$$

$$m_P := \frac{\hbar}{l_P c} = \sqrt{\frac{\hbar c}{G}} \approx 2.18 \times 10^{-5}\,\text{g} \approx \frac{1.22 \times 10^{19}\,\text{GeV}}{c^2}\,. \tag{33.6}$$

It must be emphasized that units of length, time, and mass cannot be formed out of G and c (GR) or out of $\hbar$ and c (quantum theory) alone. Similar units (with the use of the fine structure constant instead of $\hbar$) had been proposed before Planck by Stoney in 1881. In addition to the above Planck units, one can also define a *Planck charge* as follows,

$$Q_P := \sqrt{m_P l_P \frac{l_P}{t_P}} = \sqrt{G} m_P = \sqrt{\hbar c}\,, \tag{33.7}$$

which is independent of G. One can see that the elementary electric charge is $e = \sqrt{\alpha} Q_P \approx 0.085 Q_P$, with α as the fine structure constant. For two particles with electric charge Q_P and mass m_P, the Coulombian repulsion exactly compensates the Newtonian attraction.

For the study of structures in the Universe, the Planck scale is usually irrelevant. The reason is that scales of astrophysical relevance are typically connected with the *fine structure constant of gravity* defined by

$$\alpha_g := \frac{G m_{pr}^2}{\hbar c} = \left(\frac{m_{pr}}{m_P}\right)^2 \approx 5.91 \times 10^{-39}\,, \tag{33.8}$$

where m_{pr} denotes the proton mass. For example, the Chandrasekhar mass M_C (which gives the mass scale for main sequence stars) is approximately given by $M_C \approx \alpha_g^{-3/2} m_{pr} \approx 1.8 M_\odot$, which is much bigger than the Planck mass.

As far as Planck scale effects in accelerators are concerned, one would have to increase the large hadron collider (LHC) up to the size of the Milky Way, in order to create particles with masses given by (33.6). Tests of quantum gravity must thus come in a different way.

In the rest of this section, we discuss briefly the limits where gravity is treated as an external field, in which the quantum objects act dynamically. Section 33.2 is then devoted to the main approaches of a full quantum theory of gravity. Recent general introductions and reviews include, besides [33.1], [33.3–8].

33.1.2 Quantum Mechanics in an External Gravitational Field

The first encounter of quantum theory with gravity is quantum mechanics with an external Newtonian gravitational field. In the classic experiment performed in 1975 by Colella, Overhauser, and Werner (the *COW* experiment), a neutron was brought into a superposition in which the components run at different heights in the gravitational field of the Earth (Fig. 33.1). The two components experience a different phase shift causing a characteristic interference pattern when recombined in a detector.

In the experiment, the change in the interference pattern with respect to the angle θ is measured. The gravitational phase shift is calculated to read

$$\Delta\beta_g \approx \frac{m_i m_g g \lambda A \sin\theta}{2\pi\hbar^2}\,, \tag{33.9}$$

where A is the area of the parallelogram ABCD, g is the gravitational acceleration, and λ the de Broglie wave-

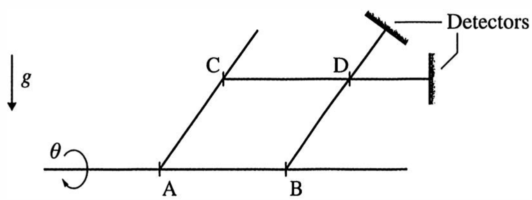


Fig. 33.1 Schematic description of the *COW* experiment for neutron interferometry in the gravitational field of the Earth (after [33.1])

length of the neutron. This expression was confirmed in the experiment with 1% accuracy.

It is interesting to see that the result for the phase shift depends on the *product* of inertial mass m_i and gravitational mass m_g. This is different from classical physics, where both quantities appear as a ratio. Nevertheless, the validity of the weak equivalence principle ($m_i = m_g$) was confirmed with an accuracy of 10^{-7}. Another important application of neutron interferometry is the observation of neutron eigenstates in the Newtonian potential (as approximated by a linear potential).

The result (33.9) can be obtained from the Schrödinger equation with a Newtonian potential. A more fundamental description employs the Dirac equation in order to take into account the spin of the neutron. By a Foldy–Wouthuysen transformation one can obtain a nonrelativistic approximation in the form (with β being equivalent to the Dirac gamma matrix γ^0),

$$i\hbar \frac{\partial \psi}{\partial t} = H_{\mathrm{FW}} \psi \,, \tag{33.10}$$

where

$$\begin{aligned} H_{\mathrm{FW}} = {} & \beta mc^2 + \frac{\beta}{2m}\mathbf{p}^2 - \frac{\beta}{8m^3c^2}\mathbf{p}^4 + \beta m(\mathbf{gx}) \\ & - \boldsymbol{\omega}(\mathbf{L}+\mathbf{S}) + \frac{\beta}{2m}\mathbf{p}\frac{\mathbf{gx}}{c^2}\mathbf{p} + \frac{\beta\hbar}{4mc^2}\boldsymbol{\Sigma}(\mathbf{g}\times\mathbf{p}) \\ & + \mathcal{O}\left(\frac{1}{c^3}\right) . \end{aligned} \tag{33.11}$$

Here, $\boldsymbol{\Sigma}$ is three spin matrices which in a convenient representation read $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma},\boldsymbol{\sigma})$, where $\boldsymbol{\sigma}$ are the Pauli matrices. The fourth term in the Hamiltonian H_{FW} is the one that leads to (33.9). The next term describes the coupling of the neutron's angular momentum and spin to the angular velocity of the Earth. The first of these couplings (leading to the Sagnac effect) has been clearly seen in the COW experiment.

If derived in this way, the weak equivalence principle is already implemented, because there is only one mass m in the Dirac equation. A fundamental derivation should start from quantum field theory, from which the Dirac equation follows in the one-particle limit only.

So far, the limit of Newtonian gravity is fully sufficient in order to describe existing experiments. A recent suggestion to include the gravitational time dilation of GR was made in [33.9]. The proper time for the upper path in Fig. 33.1 is larger than the proper time for the lower path. If the difference is sufficiently large, it may lead to a suppression of the interference pattern because *which-way information* is then available. If seen experimentally, this would probe the geometric nature of spacetime in a quantum setting.

33.1.3 Quantum Field Theory in an External Gravitational Field

The next level in the relation between quantum theory and gravity is quantum field theory in an external gravitational field. Here, no observations are available so far, although definite predictions exist. The most famous one is Hawking radiation. As Hawking found out in 1974, black holes are not really black when quantum theory is taken into account. They behave like a thermodynamical system and emit thermal radiation with a temperature proportional to $\hbar$. Explicitly, the Hawking temperature reads

$$T_{\mathrm{BH}} = \frac{\hbar\kappa}{2\pi k_{\mathrm{B}} c} \,, \tag{33.12}$$

where κ is the surface gravity of a stationary black hole, which by the no-hair theorem is uniquely characterized by its mass M, its angular momentum J, and (if present) its electric charge q. In the particular case of the spherically symmetric Schwarzschild black hole, one has $\kappa = c^4/4GM = GM/R_{\mathrm{S}}^2$, were $R_{\mathrm{S}} = 2GM/c^2$ is the Schwarzschild radius, and therefore,

$$T_{\mathrm{BH}} = \frac{\hbar c^3}{8\pi k_{\mathrm{B}} GM} \approx 6.17\times 10^{-8}\left(\frac{M_\odot}{M}\right)\mathrm{K} \,. \tag{33.13}$$

One recognizes that the black hole becomes hotter by emission ofradiation. This is because the mass is in the denominator, and the mass decreases when the emission takes place. This behavior is in contrast to the behavior of ordinary thermodynamical systems. It is, in fact, typical for gravitational systems [33.10]. The emission leads to a finite lifetime for the evaporating black hole. To understand the final phase of the evaporation, one must go beyond Hawking's approximation and apply a full theory of quantum gravity. Observational tests can only come from primordial black holes or, in the case that particular theories with higher dimensions hold, from black holes produced in colliders. Primordial black holes can be identified through their characteristic emission of γ-radiation; so far, nothing has been seen.

Since black holes behave thermodynamically, they also possess an entropy. It is given by the expression

$$S_{\mathrm{BH}} = \frac{k_{\mathrm{B}} A}{4 l_{\mathrm{P}}^2} , \tag{33.14}$$

where A is the surface of the event horizon. It is called the Bekenstein–Hawking entropy. For the special case of Schwarzschild black hole, it reads

$$S_{\mathrm{BH}} = \frac{k_{\mathrm{B}} \pi R_{\mathrm{S}}^2}{G\hbar} \approx 1.07 \times 10^{77} k_{\mathrm{B}} \left(\frac{M}{M_{\odot}}\right)^2 . \tag{33.15}$$

A major question in any theory of quantum gravity is the microscopic derivation of the area law for the entropy by counting appropriate quantum states.

There exists an analogous effect to (33.12) in flat spacetime. The concepts of vacuum and particles are usually introduced with respect to inertial observers. These notions are no longer unique if accelerated motion is considered. As Unruh found out in 1976, an observer that moves through the Minkowski vacuum with a uniform linear acceleration experiences this vacuum as filled with thermally distributed particles. The temperature is given by the *Davies–Unruh temperature*

$$T_{\mathrm{DU}} = \frac{\hbar a}{2\pi k_{\mathrm{B}} c} \approx 4.05 \times 10^{-23}\, a \left[\frac{\mathrm{cm}}{\mathrm{s}^2}\right] \mathrm{K} . \tag{33.16}$$

Although this seems to be unobservationally small, accelerations in particle detectors can, in principle, become high enough to observe this effect. The point is that the acceleration should occur in linear motion in order to see (33.16). Experiments with high-power lasers are in preparation.

The Hawking effect occurs at the level of quantum field theory on an external curved spacetime. Here, the *matter part* of the Einstein equations (33.2) consists of quantum fields. Yet, how can this be properly described? One cannot formulate an equation with operators in Hilbert space on the right-hand side and classical fields on the left-hand side. An ad hoc modification of (33.2) in order to cope with this situation is the *semiclassical Einstein equation*. In it, the energy–momentum tensor is replaced by the expectation value of the energy–momentum operator with respect to the quantum state in question,

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4} \langle \Psi | \hat{T}_{\mu\nu} | \Psi \rangle . \tag{33.17}$$

The range of validity of this equation should be understood from a full quantum theory of gravity. In the following, we shall briefly discuss the main current approaches for such a theory.

33.2 Main Approaches to Quantum Gravity

One can roughly distinguish between two classes of approaches. In the first, one starts from a given classical theory of gravity (usually, but not exclusively, GR) and then employs heuristic quantization rules to find a quantum theory of the gravitational field. This procedure does not yet by itself entail a unification of interactions. In the second, one first attempts to construct a unified quantum theory of all interactions and then tries to recover quantum gravity in the limit where the various interactions become distinguishable.

As for the first class, we shall in the following consider only the quantization of GR and distinguish between covariant and canonical approaches. In the covariant approaches, the four-dimensional covariance plays a guiding role in the formalism. This is most clearly seen by using path integrals and DeWitt's background field method. In the canonical approaches, one develops a Hamiltonian formalism at the classical level and then performs the quantization by imposing commutator relations for the canonical variables.

As for the second class, the main representative is string theory. It is supposed to be a quantum theory (beyond quantum field theory) of all interactions, often called *theory of everything*. There are other approaches that start from fundamental discrete structures and attempt to construct a unified quantum theory from them, cf. [33.7]. Among them are the theory of causal sets and group field theory. In the following, we set $c = 1$ in most expressions.

33.2.1 Covariant Quantum Gravity

The oldest covariant approach is perturbation theory around a given background, which is usually taken flat. Let us, therefore, first have a brief look at the treatment of weak gravitational waves in GR. We take for

the background the flat Minkowski spacetime with the standard metric $\eta_{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1)$ and call the perturbation $f_{\mu\nu}$. We then write

$$g_{\mu\nu} = \eta_{\mu\nu} + f_{\mu\nu} \,. \tag{33.18}$$

Instead of $f_{\mu\nu}$, it is often useful to use the following combination,

$$\bar{f}_{\mu\nu} := f_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu} f^{\rho}_{\ \rho} \,. \tag{33.19}$$

The linearized Einstein equations then assume the simple form

$$\Box \bar{f}_{\mu\nu} = -16\pi G T_{\mu\nu} \,, \tag{33.20}$$

if the *harmonic gauge condition* $\partial_\nu \bar{f}_\mu^{\ \nu} = 0$ is used.

Quantization is performed by turning the perturbation $f_{\mu\nu}$ into an operator. If the above weak gravitational waves are quantized, one arrives at linearized quantum gravity, which was first discussed in this way by Bronstein in 1936 and elaborated on by Gupta in 1952. At the linearized level, there is a close analogy with electrodynamics. While the mediator of the free electromagnetic field is the photon, a massless particle of spin one, the mediator of the linearized gravitational field is the *graviton*, a massless particle of spin two. Empirically, the mass of the graviton is currently limited by 10^{-29} eV.

Formally, one starts from a superposition of plane waves,

$$f_{\mu\nu}(x) = \sum_{\sigma=\pm 2} \int \frac{\mathrm{d}^3 k}{\sqrt{2|\mathbf{k}|}} \times \left[a(\mathbf{k},\sigma) e_{\mu\nu}(\mathbf{k},\sigma) \mathrm{e}^{\mathrm{i}kx} + a^\dagger(\mathbf{k},\sigma) e^*_{\mu\nu}(\mathbf{k},\sigma) \mathrm{e}^{-\mathrm{i}kx} \right] , \tag{33.21}$$

and turns the amplitudes into operators satisfying

$$[a(\mathbf{k},\sigma), a^\dagger(\mathbf{k}',\sigma')] = \delta_{\sigma\sigma'} \delta(\mathbf{k}-\mathbf{k}') \,, \tag{33.22}$$

with all other commutators vanishing.

Already at this level, effects of quantum gravity can be discussed. One can, for example, calculate the transition rate from the 3-D level to the 1s level in the hydrogen atom due to the emission of a graviton. One obtains

$$\Gamma_{\mathrm{g}} = \frac{G m_{\mathrm{e}}^3 c \alpha^6}{360 \hbar^2} \approx 5.7 \times 10^{-40}\,\mathrm{s}^{-1} \,, \tag{33.23}$$

which corresponds to a lifetime of

$$\tau_{\mathrm{g}} \approx 5.6 \times 10^{31}\,\mathrm{yr} \,. \tag{33.24}$$

This seems too long to be observable, although it is of the same order as the value for the proton lifetime predicted by some unified theories, which has been excluded experimentally.

Another observable effect of linear quantum gravity may, in fact, lie around the corner. According to the inflationary scenario of cosmology, gravitons were produced in the early Universe. These gravitons would exhibit themselves as a tensor contribution to the CMB anisotropy spectrum. A detection of this contribution has been announced by the BICEP2 experiment in March 2014 [33.11].

In going beyond the linear level, the most straightforward way is to employ a path integral quantization [33.12]. The quantum gravitational path integral was first formulated by Misner in 1957 and formally reads

$$Z[g] = \int \mathcal{D} g_{\mu\nu}(x) \mathrm{e}^{\mathrm{i}S[g_{\mu\nu}(x)]} \,, \tag{33.25}$$

where the sum runs over all metrics on a four-dimensional manifold $\mathcal{M}$ quotiented by the diffeomorphism group Diff$\mathcal{M}$. In addition, one would like to perform a sum over all topologies, but this is not possible in full generality, since four-manifolds are not classifiable. Considerable care must be taken in the treatment of the integration measure. This includes the application of the Faddeev–Popov procedure [33.1].

In order to use the path integral for the derivation of Feynman rules, one generalizes the ansatz (33.18) to

$$g_{\mu\nu} = \bar{g}_{\mu\nu} + \sqrt{32\pi G} f_{\mu\nu} \,, \tag{33.26}$$

where $\bar{g}_{\mu\nu}$ denotes the background field with respect to which (four-dimensional) covariance will be implemented in the formalism, and $f_{\mu\nu}$ denotes the quantized field, which has here the dimension of a mass, and which is here taken to all orders. The covariance with respect to the background metric means that no particular background is distinguished; that is, *background independence* is implemented into the formalism (this method is also referred to as *DeWitt's background field method*, a somewhat misleading terminology).

In contrast to linear quantum gravity, the formalism now allows for graviton interactions. As is usual in quantum field theory, divergences occur. A major problem arises in the treatment of these divergences. In

contrast to quantum electrodynamics (QED) and Yang–Mills theories, perturbative quantum gravity turns out to be nonrenormalizable. This means that at each order of the perturbation theory new types of divergences emerge, which must be absorbed by new parameters to be introduced into the action. This leaves one with infinitely many parameters (each of which must be determined experimentally), rendering, it seems, the theory meaningless. The formal reason is that the expansion parameter – the gravitational constant appearing in (33.26) – possesses a negative mass dimension.

For pure gravity, the divergences are absent on-shell at the one-loop level. They occur, however, from the two-loop order on. Using dimensional regularization, the divergent part in the two-loop Lagrangian reads

$$\mathcal{L}^{(\mathrm{div})}_{2-\mathrm{loop}} = \frac{209\hbar^2}{2880}\frac{32\pi G}{(16\pi^2)^2\epsilon}\sqrt{-\bar{g}}\bar{R}^{\alpha\beta}_{\ \ \gamma\delta}\bar{R}^{\gamma\delta}_{\ \ \mu\nu}\bar{R}^{\mu\nu}_{\ \ \alpha\beta}\,, \tag{33.27}$$

where $\epsilon = 4-D$, with D being the number of spacetime dimensions.

Does this nonrenormalizability really render the perturbative approach meaningless? Not necessarily. In the limit of low energies, the arbitrariness coming from the infinitely many renormalization parameters disappears, and definite results can be calculated in the ensuing effective theory. One example discussed by *Bjerrum-Bohr* et al. [33.13] in 2003 is the quantum gravitational correction to the Newtonian potential between two masses m_1 and m_2, for which they find

$$V(r) = -\frac{Gm_1m_2}{r}\times\left(1+3\frac{G(m_1+m_2)}{rc^2}+\frac{41}{10\pi}\frac{G\hbar}{r^2c^3}+\mathcal{O}(G^2)\right). \tag{33.28}$$

(The first correction term, which does not contain $\hbar$, is a well-known correction from classical GR.) The quantum gravitational correction term is too small to be seen in laboratory experiments; the notable point is that a definite term can be calculated.

A modification of GR that leads to a perturbative theory of quantum gravity that is renormalizable (and also unitary) is Hořava–Lifshitz gravity [33.14]. This comes, however, at the price of violating Lorentz invariance (and thus the equivalence principle) at high energies. Its status is thus open.

Using path integral formalism, one can derive the semiclassical Einstein equations (33.17) at the one-loop level. One finds (neglecting, here, the cosmological constant) [33.1]

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = 8\pi G\left(\langle T_{\mu\nu}\rangle + \langle t_{\mu\nu}\rangle\right), \tag{33.29}$$

where the left-hand side is evaluated for the mean metric $\langle g_{\mu\nu}\rangle$. Compared with (33.17), there is also a contribution from the gravitons through their energy–momentum tensor $t_{\mu\nu}$. This demonstrates that at the one-loop order (the limit of quantum theory in curved spacetime), the gravitons are as important as the matter fields.

The gravitational path integral (33.25) is either evaluated in a semiclassical (saddle point) approximation or numerically. For a numerical evaluation, either Regge calculus [33.12] or causal dynamical triangulation (Chap. 34) is used. For this purpose, spacetime is decomposed into a set of simplices. In Regge calculus, a Euclidean formulation is employed, and the edge lengths of the simplices are treated as dynamical entities. In causal dynamical triangulation, a Lorentzian formulation is used, and the edge lengths are kept fixed, while the sum in the path integral is performed over all possible manifold-gluing of simplices. Here, one result is a phase transition from a four-dimensional spacetime at large scales to a two-dimensional one at small scales.

A transition to a two-dimensional spacetime at small scales can also be seen in the approach of asymptotic safety [33.15]. In quantum field theory, coupling constants are actually energy-dependent due to renormalization. This dependence is described by renormalization group equations. A theory is called asymptotically safe if at least one of the coupling constants approaches a nontrivial (i. e., nonvanishing) fixed point at energy $E \to \infty$. (If the coupling constants approach a vanishing fixed point, the theory is called asymptotically free; this happens, for example, in quantum chromodynamics (QCD) and perhaps also QED.) Detailed calculations indicate that quantum gravity may, in fact, be asymptotically safe [33.15].

So far, only the quantization of GR has been addressed. The situation may change if supersymmetry is added, leading to a theory of supergravity (SUGRA), in which fermions and bosons appear on equal footing. Recent work has found indications that a quantum field theory based on $N = 8$ SUGRA (this corresponds to the maximal number of supersymmetry generators) may be perturbatively finite. More precisely, it was shown

that the theory is finite up to four loops, and arguments were presented that finiteness also holds at five and six loops [33.16]. This could mean that a (so far unknown) symmetry guarantees the finiteness of $N = 8$ SUGRA at all orders. If this were true, one would then have a constructed a finite quantum field theory for the gravitational field.

33.2.2 Canonical Approaches

In the canonical approaches to quantum gravity, one develops a Hamiltonian formulation at the classical level and then imposes commutator relations for the canonical variables. For this purpose, a $3+1$ decomposition of spacetime is performed (for which the assumption is needed that spacetime is globally hyperbolic) [33.1]. The four-dimensional metric is thereby decomposed into a three-dimensional metric, one lapse function, and a shift vector. In this procedure, one arrives at constraints on the canonical variables. Such constraints arise in any theory that is classically invariant under redundancy transformations. In GR, these are the spacetime diffeomorphisms. One thus has four constraints (per space point). One is the Hamiltonian constraint, which generates hypersurface deformations (many-fingered time evolution). The three others are the momentum or diffeomorphism constraints, which generate three-dimensional coordinate transformations. If a tetrad formulation of GR is used, one has in addition three *Gauss constraints*, which generate triad rotations; they play a role in loop quantum gravity. Classically, the constraints obey a closed (but not Lie) algebra.

To connect the redundancy transformations at the canonical level with spacetime diffeomorphisms, one must note that the corresponding redundancy generator is a sum of these constraints plus additional constraints that are related with lapse function and shift vector. The Hamiltonian constraint, although part of this generator, can nevertheless by itself generate physical motion [33.17].

In the quantum theory, the constraints are transformed into quantum constraints that act on physically allowed wave functions. Depending on the variables used, one distinguishes between quantum geometrodynamics and loop quantum gravity. An important consistency requirement in both approaches is the off-shell closure of the quantum constraint algebra (also known as Dirac consistency). Whether this can be achieved or not, is presently open. Closure of the algebra means absence of central terms. In string theory (see below), these terms are, however, crucial; they yield, for example, the critical number of dimensions in which the string can propagate. If such central terms are also needed in canonical quantum gravity, the standard formalism presented below must be modified.

Quantum Geometrodynamics. In the geometrodynamical version of canonical gravity, the canonical variables are the three-metric $h_{ab}(x)$ and its conjugate momentum $p^{cd}(y)$, which is a linear combination of the second fundamental form. Upon quantization, one turns these variables into operators and imposes the formal commutator relations

$$[\hat{h}_{ab}(x), \hat{p}^{cd}(y)] = \mathrm{i}\hbar\delta^c_{(a}\delta^d_{b)}\delta(x,y) . \tag{33.30}$$

They are formal because they do not implement the requirement that the determinant h of the three-metric be positive. In this way, one arrives at the quantum constraints

$$\mathcal{H}_\perp \Psi = 0, \tag{33.31}$$

$$\mathcal{H}_a \Psi = 0 . \tag{33.32}$$

The first equation is called the Wheeler–DeWitt equation, the three other equations are called quantum momentum or diffeomorphism constraints. Equations (33.32) guarantee that the wave functional is invariant under three-dimensional coordinate transformations. In the vacuum case, these equations read explicitly

$$\hat{\mathcal{H}}_\perp \Psi := \left(-16\pi G\hbar^2 G_{abcd}\frac{\delta^2}{\delta h_{ab}\delta h_{cd}} - \frac{\sqrt{h}}{16\pi G}\left({}^{(3)}R - 2\Lambda\right)\right)\Psi = 0 , \tag{33.33}$$

$$\hat{\mathcal{H}}_a \Psi := -2D_b h_{ac}\frac{\hbar}{\mathrm{i}}\frac{\delta\Psi}{\delta h_{bc}} = 0 . \tag{33.34}$$

Here, D_b denotes the three-dimensional covariant derivative, G_{abcd} is the DeWitt metric (an ultralocal function of the three-metric), and ${}^{(3)}R$ is the three-dimensional Ricci scalar. A detailed discussion of these equations can be found in [33.18, 19] and [33.1]. Because of the notorious factor-ordering problem, (33.33) possesses formal meaning only.

A most surprising feature of these equations is the absence of spacetime – only the three-metric remains in the formalism. Quantum gravity in this form thus seems to be fundamentally timeless. Upon second thought, however, this feature is less surprising [33.20]. Classical spacetime is the analog of a particle trajectory

in classical mechanics. After quantization, the particle trajectory disappears, and so does spacetime. The timelessness is thus a natural consequence of a quantum theory of gravity in the absence of an absolute external time.

There is a close connection with the path integral formalism discussed above. One can show, at least at a formal level, that the full path integral of quantum gravity satisfies the quantum constraints (33.31) and (33.32). The situation is thus analogous to quantum mechanics, where the path integral satisfies the Schrödinger equation. The factor-ordering problem in the Wheeler–DeWitt equation corresponds to the measure ambiguity in the path integral.

The treatment of functional differential equations such as (33.31) and (33.32) is notoriously difficult. In this sense, the situation is not better (but also not worse) than the situation for the functional Schrödinger equation in, for example, QCD. One possibility to deal with the full set of equations is the use of lattice methods [33.21]. Another possibility is to employ approximations. Among the latter is the Wentzel–Kramers–Brillouin (WKB) approximation, which corresponds to the saddle-point approximation for the path integral. One makes the ansatz

$$\Psi[h_{ab}] = C[h_{ab}] \exp\left(\frac{\mathrm{i}}{\hbar} S[h_{ab}]\right) , \qquad (33.35)$$

where $C[h_{ab}]$ is a *slowly varying amplitude* and $S[h_{ab}]$ is a *rapidly varying phase* (an *eikonal* as in geometrical optics).

Yet another approximation scheme makes use of the fact that the energy scales for nongravitational fields are usually much smaller than the Planck mass. One can then employ a Born–Oppenheimer type of approximation, which is known from molecular physics, where the different scales are the masses of the nuclei and the electrons. In this way, one can derive the limit of quantum theory in an external background in the form of a (functional) Schrödinger equation for matter fields propagating in a spacetime described by Einstein's equations. The (many-fingered) time in this equation arises from the state of the gravitational field as a *WKB time*. At higher orders of the Born–Oppenheimer scheme, one can derive quantum gravitational corrections to this Schrödinger equation. These correction terms lead, in principle, to observable effects. One example is the modification of the CMB anisotropy spectrum at large scales, although this modification seems too small to be presently observable [33.22, 23].

Another type of approximation is the quantization of classically reduced models. Let us briefly review two of their classes. The first addresses the reduction to spherically symmetric models [33.1, 24]. This yields models that include the quantum Schwarzschild black hole and the quantum version of the Lemaître–Tolman–Bondi model (classically describing the collapse of a spherically symmetric dust cloud). In the latter, one can also attempt to reproduce the entropy (33.14) from the quantum states. This leads to $S_{\mathrm{BH}} \propto A$, but not with the correct proportionality factor.

The second class is quantum cosmology. Here, one performs a classical reduction to homogeneous models, mostly Friedmann–Lemaître models, but also models for anisotropic universes [33.1, 25]. The Wheeler–DeWitt equation is well defined in those cases and can be solved at least numerically, if not analytically. For some models and for some boundary conditions, one can show that the solutions avoid the classical singularities. This happens, for example, if the wave function turns out to be zero in the region where the classical singularities lurk. A general statement cannot be made so far. In quantum cosmology, one can study conceptual questions as well as the potential observational effects. One can even attempt to understand the origin of irreversibility [33.10]. Classical properties emerge from the quantum universe in an approximate way through decoherence, a concept that is well understood from quantum mechanics [33.26]. One can also study the supersymmetric extension of quantum cosmology [33.27].

Loop Quantum Gravity. Loop quantum gravity is a canonical approach that uses a set of canonical variables different from geometrodynamics. It uses variables that are conceptually closer to variables familiar from Yang–Mills theories. A detailed treatment can be found in [33.28–32].

The starting point is *Ashtekar's new variables*, introduced by Ashtekar in 1986. The role of the momentum variable is played by the densitized triad (dreibein)

$$E_i^a(x) := \sqrt{h}(x) e_i^a(x) , \qquad (33.36)$$

while the momentum variable is the connection

$$GA_a^i(x) = \Gamma_a^i(x) + \beta K_a^i(x) . \qquad (33.37)$$

Here, a (i) denotes a space index (internal index), $\Gamma_a^i(x)$ is the spin connection, and $K_a^i(x)$ is related to the second fundamental form. The parameter β is called the

Barbero–Immirzi parameter and can assume any nonvanishing real value; it represents a new freedom for the quantum theory.

The canonical variables obey

$$\{A_a^i(x), E_j^b(y)\} = 8\pi\beta\delta_j^i\delta_a^b\delta(x,y)\,. \tag{33.38}$$

There are again the Hamiltonian and momentum constraints known from geometrodynamics, but rewritten in terms of the new variables. Because of the use of the triad, one encounters a new constraint called the Gauss constraint, which expresses the local freedom to rotate these triads.

The loop variables are constructed from these variables in a nonlocal manner. The new connection variable is the holonomy $U[A,\alpha]$, which is a path-ordered exponential of G times the integral over the connection around a loop α. In the quantum theory, it acts on wave functionals as

$$\hat{U}[A,\alpha]\Psi_S[A] = U[A,\alpha]\Psi_S[A]\,. \tag{33.39}$$

The new momentum variable is the flux of the densitized triad through a two-dimensional surface S bounded by the loop. Its operator version reads

$$\hat{E}_i[S] := -8\pi\beta\hbar\mathrm{i}\int_S \mathrm{d}\sigma^1\,\mathrm{d}\sigma^2\, n_a(\boldsymbol{\sigma})\frac{\delta}{\delta A_a^i[\mathbf{x}(\boldsymbol{\sigma})]}\,, \tag{33.40}$$

where the embedding of the surface is given by $(\sigma^1,\sigma^2) \equiv \boldsymbol{\sigma} \mapsto x^a(\sigma^1,\sigma^2)$. The variables obey the commutation relations

$$\left[\hat{U}[A,\alpha], \hat{E}_i[S]\right] = \mathrm{i}l_{\mathrm{P}}^2\beta\iota(\alpha,S)U[\alpha_1,A]\tau_i U[\alpha_2,A]\,,$$

where $\iota(\alpha,S) = \pm 1, 0$ is the *intersection number*, which depends on the orientation of α and S. For these commutation relations, one can prove a theorem that is analogous to the Stone–von Neumann theorem in quantum mechanics: the holonomy flux representation is essentially unique. This gives rise to a unique Hilbert space structure at the kinematical level (before the constraints are imposed). An important feature of this Hilbert space is its nonseparable character, that is, it does not admit a countable basis. A convenient basis is the spin network basis, which consists of graphs with spins attached to it.

The spin network structure suggests that space may be discrete at small scales. One can, in fact, construct a self-adjoint area operator which has a discrete spectrum on the kinematical Hilbert space,

$$\begin{aligned}\hat{\mathcal{A}}[S]\Psi_S[A] &= 8\pi\beta l_{\mathrm{P}}^2\sum_{\mathrm{P}\in S\cap S}\sqrt{j_{\mathrm{P}}(j_{\mathrm{P}}+1)}\Psi_S[A]\\ &=: A[S]\Psi_S[A]\,.\end{aligned} \tag{33.41}$$

Here, P denotes the intersection points between the spin network S and the surface S, and j_{P} can assume integer and half-integer values (arising from the use of the group SU(2) for the triads). There thus exists a minimal *quantum of action* of the order of β times the Planck length squared. Whether this discrete structure is preserved at the dynamical level (after the constraints are imposed) is far from clear. As in quantum geometrodynamics, a full understanding of the quantum constraints is elusive; this holds, in particular, for the Hamiltonian constraint (the loop version of the Wheeler–DeWitt equation). Also elusive is a precise formulation of the semiclassical limit.

There also exists a covariant version of loop quantum gravity. It corresponds to a path integral formulation, through which the spin networks are evolved *in time*. It is referred to as the spin foam approach [33.33].

It has been attempted to apply loop quantum gravity to a microscopic derivation of the black hole entropy (33.14). Treating the microstates as distinguishable, it was indeed found that $S_{\mathrm{BH}} \propto A$, with a proportionality factor that depends on β. The result coincides with (33.14) for a very peculiar value of β; whether this value has any significance, is open.

As in quantum geometrodynamics, one can also apply loop quantum gravity to cosmology, resulting in loop quantum cosmology [33.34–36]. The Wheeler–DeWitt equation is replaced by a difference equation. This equation also provides the means for singularity avoidance. The same conclusion results from a bounce predicted from an effective Friedmann equation; this bounce gives a lower bound for the size of the universe, preventing the occurrence of the big bang. Loop quantum cosmology has also been applied to the CMB anisotropy spectrum, for which it predicts an enhancement at large scales [33.23].

33.2.3 String Theory

String theory is fundamentally different from the approaches discussed above. It is not a quantization of GR or any other classical theory of gravity. It has

the ambition to be a fundamental quantum theory in which all interactions are unified. Gravity, as well as the other known interactions, only emerge in an appropriate limit. Strings are one-dimensional objects characterized by a dimensionful parameter α' or the string length $l_s = \sqrt{2\alpha'\hbar}$ constructed from it. In spacetime, it forms a two-dimensional surface, the worldsheet. Closer inspection of the theory also exhibits the presence of higher-dimensional objects called D-branes, which are as important as the strings themselves. Extensive treatments of string theory include [33.37, 38]; a recent assessment from a conceptual point of view is given in [33.39].

String theory necessarily contains gravity, because the graviton appears as an excitation of closed strings. It is through this appearance that a connection to covariant quantum gravity discussed above can be made. String theory also includes gauge theories, since the corresponding gauge bosons are found in the spectrum. It also needs the presence of supersymmetry for a consistent formulation. Fermions are thus an important ingredient of string theory. One recognizes that gravity, other fields, and matter appear on the same footing.

Because of reparametrization invariance on the worldsheet, string theory also possesses constraint equations. The constraints do not, however, close, but contain a central term on the right-hand side. This corresponds to the presence of an anomaly (connected with Weyl transformations). The vanishing of this anomaly can be achieved if ghost fields are added that gain a central term which cancels the original one. The important point is that this works only in a particular number D of dimensions: $D = 26$ for the bosonic string, and $D = 10$ for the superstring. The presence of higher spacetime dimensions is an essential ingredient of string theory.

Let us consider, for simplicity, the bosonic string. Its quantization is usually performed through the Euclidean path integral

$$Z = \int \mathcal{D}X\mathcal{D}h e^{-S_P}\,, \tag{33.42}$$

where X and h are a shorthand for the embedding variables and the worldsheet metric, respectively. The action in the exponent is the *Polyakov action*, which is an action defined on the worldsheet. Besides the dynamical variables X and h, it contains various background fields on spacetime, among them the metric of the embedding space and a scalar field called dilaton. It is obvious that this formulation is not background independent.

If the string propagates in a curved spacetime with the metric $g_{\mu\nu}$, the demand for the absence of a Weyl anomaly leads to consistency equations that correspond (up to terms of order α') to the Einstein equations for the background fields. These equations can be obtained from an effective action of the form

$$\begin{aligned} S_{\text{eff}} \propto &\int \mathrm{d}^D x\sqrt{-g}\,e^{-2\Phi} \\ &\times \left(R - \frac{2(D-26)}{3\alpha'} - \frac{1}{12}H_{\mu\nu\rho}H^{\mu\nu\rho}\right. \\ &\left. + 4\nabla_\mu\Phi\nabla^\mu\Phi + \mathcal{O}(\alpha')\right), \end{aligned} \tag{33.43}$$

where Φ is the dilaton, R the Ricci scalar corresponding to $g_{\mu\nu}$, and $H_{\mu\nu\rho}$ the field strength associated with an antisymmetric tensor field (which in $D = 4$ would be the axion). This is the second connection of string theory with gravity, after the appearance of the graviton as a string excitation.

The connection with the covariant perturbation theory of quantum gravity is then performed by making the ansatz (33.26) and inserting this into the effective action (33.43). Comparison is then made through scattering amplitudes. For example, the amplitude for graviton–graviton scattering from the scattering of strings at tree level coincides with the corresponding amplitude in the covariant perturbation theory. In this way, one can connect the gravitational constant with the string length.

The finiteness of the string length has consequences for the microstructure of spacetime. As gedanken experiments show, there seems to be a minimal length, at least in an operational sense. Veneziano has found a generalized uncertainty relation of the form

$$\Delta x \geq \frac{\hbar}{\Delta p} + \frac{l_s^2}{\hbar}\Delta p\,.$$

This relation can be heuristically extended to a general class of uncertainty relations, also called the generalized uncertainty principle. From such relations, one can calculate corrections to effects such as the Lamb shift, which have potential observational significance. (So far, nothing has been seen.)

Concerning black holes, it is possible to give a microscopic derivation of (33.14) for extremal and near-extremal *stringy* black holes. The treatment of the ordinary Schwarzschild black hole remains elusive. The derivation is somewhat indirect, though. One exploits the existence of *BPS states* for which the mass is fixed in terms of their charges, and whose spectrum is preserved on the transition from the weak to the strong

coupling limit of string theory. In the weak coupling limit, a BPS state can describe a bound state of D-branes, whose entropy can easily be calculated by string methods. In the strong coupling limit, the state can describe an extremal black hole. It turns out that the entropy calculated in the weak coupling limit exactly coincides with the expression (33.14) for the extremal black hole.

For extremal black holes, the Hawking temperature vanishes. It is possible, however, to generalize the calculation of the entropy to near extremal black holes for which there is Hawking radiation. The radiation corresponds here to the emission of closed strings from D-branes. If the D-brane state is traced out in the full quantum state, the radiation is described by a mixed thermal state. The full quantum state evolves unitarily, which is why no *information loss problem* is present, at least at the semiclassical level where the calculations are performed. An understanding of the final evaporation phase is as elusive in string theory as it is in the other approaches.

The black hole entropy (33.14) involves the area of the event horizon, not the volume inside. This gives rise to the idea of the *holographic principle* according to which the information (or missing information) for a gravitating system is located on the boundary of a spatial region. This principle seems to be realized in string theory in the form of the AdS-CFT correspondence [33.40]. This correspondence states that nonperturbative string theory in a background spacetime that is asymptotically anti-de Sitter (AdS) is dual to a conformal field theory (CFT) defined in a flat spacetime with one lower dimension. It associates fields in string theory with operators in the CFT and compares expectation values and symmetries in the two theories; an equivalence at the level of the quantum states has not been shown. The AdS-CFT correspondence provides an almost background-independent definition of string theory because the background metric enters only through boundary conditions at infinity [33.41]. A truly background-independent formulation may be provided by string field theory, but this is an open issue.

33.3 Outlook

Where do we stand? It is fair to say that none of the above approaches has been proven to be *the* quantum theory of gravity. The main problem is the lack of empirical tests so far, although attempts to find such tests are being undertaken in various directions [33.42, 43]. Unfortunately, no sign of new physics is seen at the LHC so far.

It is, again, important to emphasize the different nature of the above approaches. The covariant and canonical approaches aim at the construction of a separate quantum theory of the gravitational field; the unification with other interactions is a secondary feature. It is thus not surprising that the *matter aspect* has been neglected there compared with the *gravity aspect*. These approaches are thus most likely effective theories only, in the same sense that QED is an effective theory. They nevertheless provide important insight and can in principle be tested by observations. The central requirement of background independence (absence of absolute structures) is implemented in these approaches.

String theory is so far the only serious candidate for a unified quantum theory of all interactions. Gravity is, like the other forces, an emergent interaction only. The many particles in Nature can, in principle, be understood from string excitations, analogously to the way the elements in the atomic table can be understood from electrons and protons. Of particular importance is the relevance of supersymmetry. Fermions are thus an indispensable ingredient of the theory. This is not seen in the other approaches. However, string theory also suffers from problems. It seems that there are more than 10^{500} possible ground states of the theory (the infamous *string landscape*). Still, it is difficult if not impossible to recover the standard model from it. It has been claimed that a selection among the many ground states in the landscape can only made by the anthropic principle, but this would go strongly against the original idea of string theory to find a unique description of Nature.

Gravity is the oldest known interaction and the one that is of immediate relevance for everyday life. It is amazing that it is also the interaction that still presents the greatest mystery.

References

33.1 C. Kiefer: *Quantum Gravity*, 3rd edn. (Oxford Univ. Press, Oxford 2012)

33.2 S.W. Hawking, R. Penrose: *The Nature of Space and Time* (Princeton Univ. Press, Princeton 1996)

33.3 H. Nicolai: Quantum Gravity: The view from particle physics, arXiv:1301.5481v1 [gr-qc] (2013)

33.4 S. Carlip: Quantum gravity; A progress report, Rep. Prog. Phys. **64**, 885–942 (2001)

33.5 G. Esposito: An introduction to quantum gravity, arXiv: 1108.3269v1 [hep-th] (2011)

33.6 R.P. Woodard: How far are we from the quantum theory of gravity?, Rep. Prog. Phys. **72**, 126002 (2009)

33.7 D. Oriti (Ed.): *Approaches to Quantum Gravity* (Cambridge Univ. Press, Cambridge 2009)

33.8 G. Calcagni, L. Papantonopoulos, G. Siopsis, N. Tsamis (Eds.): *Quantum Gravity and Quantum Cosmology*, Lecture Notes in Physics, Vol. 863 (Springer, Berlin, Heidelberg 2013)

33.9 M. Zych, F. Costa, I. Pikovski, T.C. Ralph, C. Brukner: General relativistic effects in quantum interference of photons, Class. Quantum Gravity **29**, 224010 (2012)

33.10 H.D. Zeh: *The Physical Basis of the Direction of Time*, 5th edn. (Springer, Berlin, Heidelberg 2007)

33.11 BICEP2 collaboration: Detection of B-mode polarization at degree angular scales, arXiv:1403.3985v2 [astro-ph.CO] (2014)

33.12 H. Hamber: *Quantum Gravitation: The Feynman Path Integral Approach* (Springer, Berlin, Heidelberg 2009)

33.13 N.E.J. Bjerrum-Bohr, J.F. Donoghue, B.R. Holstein: Quantum gravitational corrections to the nonrelativistic scattering potential of two masses, Phys. Rev. D **67**, 084033 (2003)

33.14 T.P. Sotiriou: Hořava–Lifshitz gravity: a status report, J. Phys. Conf. Ser. **283**, 012034 (2011)

33.15 M. Reuter, F. Saueressig: Asymptotic Safety, Fractals, and Cosmology, arXiv:1205.5431v1 [hep-th] (2012)

33.16 Z. Bern, J.J. Carrasco, L.J. Dixon, H. Johansson, R. Roiban: Amplitudes and ultraviolet behavior of $N = B$ supergravity, Fortschr. Phys. **59**, 561–578 (2011)

33.17 J. Barbour, B. Z. Foster: Constraints and gauge transformations: Dirac's theorem is not always valid, arXiv:0808.1223v1 [gr-qc] (2008)

33.18 B.S. DeWitt: Quantum theory of gravity. I. The canonical theory, Phys. Rev. **160**, 1113–1148 (1967)

33.19 J.A. Wheeler: Superspace and the nature of quantum geometrodynamcis. In: *Battelle Rencontres*, ed. by C.M. DeWitt, J.A. Wheeler (Benjamin, New York 1968) pp. 242–307

33.20 C. Kiefer: Does time exist in quantum gravity? Second prize essay of the Foundational Questions Institute essay contest on the nature of time, arXiv:0909.3767v1 [gr-qc] (2009)

33.21 H.W. Hamber, R. Toriumi, R.M. Williams: Wheeler–DeWitt Equation in $3+1$ Dimensions, arXiv:1212.3492 [hep-th] (2012)

33.22 D. Bini, G. Esposito, C. Kiefer, M. Krämer, F. Pessina: On the modification of the cosmic microwave background anisotropy spectrum from canonical quantum gravity, Phys. Rev. D **87**, 104008 (2013)

33.23 G. Calcagni: Observational effects from quantum cosmology, Ann. Phys. **525**, 323 (2013)

33.24 J.F.G. Barbero, E.J. Villaseñor: Quantization of midisuperspace models, Living Rev. Relativ. **13**, 6 (2010)

33.25 J.J. Halliwell: Introductory lectures on quantum cosmology. In: *Quantum Cosmology and Baby Universes*, ed. by S. Coleman, J.B. Hartle, T. Piran, S. Weinberg (World Scientific, Singapore 1991) pp. 159–243

33.26 E. Joos, H.D. Zeh, C. Kiefer, D. Giulini, J. Kupsch, I.-O. Stamatescu: *Decoherence and the Appearance of a Classical World in Quantum Theory*, 2nd edn. (Springer, Berlin, Heidelberg 2003)

33.27 P.V. Moniz: *Quantum Cosmology: The Supersymmetric Perspective*, Lecture Notes in Physics, Vol. 804 (Springer, Berlin, Heidelberg 2010)

33.28 C. Rovelli: *Quantum Gravity* (Cambridge Univ. Press, Cambridge 2007)

33.29 T. Thiemann: *Modern Canonical Quantum General Relativity* (Cambridge Univ. Press, Cambridge 2007)

33.30 A. Ashtekar, J. Lewandowski: Background independent quantum gravity: A status report, Class. Quantum Gravity **21**, R53–152 (2004)

33.31 H. Nicolai, K. Peeters, M. Zamaklar: Loop quantum gravity: an outside view, Class. Quantum Gravity **22**, R193–R247 (2005)

33.32 R. Gambini, J. Pullin: *A first course in loop quantum gravity* (Oxford Univ. Press, Oxford 2011)

33.33 C. Rovelli: Covariant Loop Gravity. In: *Quantum Gravity and Quantum Cosmology*, Lecture Notes in Physics, Vol. 863, ed. by G. Calcagni, L. Papantonopoulos, G. Siopsis, N. Tsamis (Springer, Berlin, Heidelberg 2013) pp. 57–66

33.34 M. Bojowald: *Quantum cosmology*, Lecture Notes in Physics, Vol. 835 (Springer, Berlin, Heidelberg 2011)

33.35 A. Ashtekar, P. Singh: Loop quantum cosmology: A status report, Class. Quantum Gravity **28**, 213001 (2011)

33.36 M. Bojowald, C. Kiefer, P.V. Moniz: Quantum cosmology for the 21st century: A debate, arXiv:1005.2471v1 [gr-qc] (2010)

33.37 M.B. Green, J.H. Schwarz, E. Witten: *Superstring Theory* (Cambridge Univ. Press, Cambridge 1987), 2 Vols.

33.38 R. Blumenhagen, D. Lüst, S. Theisen: *Basic Concepts of String Theory* (Springer, Berlin, Heidelberg 2013)

33.39 Found Phys. (Special issue): Forty years of string theory: Reflecting on the foundations, 43(1) (2013)

33.40 J. Maldacena: The gauge/gravity duality, arXiv:1106.6073v1 [hep-th] (2011)

33.41 M. Blau, S. Theisen: String theory as a theory of quantum gravity: A status report, Gen. Relativ. Gravit. **41**, 743–755 (2009)

33.42 G. Amelino-Camelia, J. Kowalski-Glikman (Eds.): *Planck Scale Effects in Astrophysics and Cosmology*, Lecture Notes in Physics, Vol. 669 (Springer, Berlin, Heidelberg 2005)

33.43 Class. Quantum Gravity (Special issue): Tests of the weak equivalence principle, 29(18) (2012)

34. Quantum Gravity via Causal Dynamical Triangulations

Jan Ambjørn, Andrzej Görlich, Jerzy Jurkiewicz, Renate Loll

Causal dynamical triangulations (CDT) represent a lattice regularization of the sum over spacetime histories, providing us with a nonperturbative formulation of quantum gravity. The ultraviolet fixed points of the lattice theory can be used to define a continuum quantum field theory, potentially making contact with quantum gravity defined via asymptotic safety. We describe the formalism of CDT, its phase diagram, and the *quantum geometries* emerging from it. We also argue that the formalism should be able to describe a more general class of quantum-gravitational models of Hořava–Lifshitz type.

34.1 Asymptotic Safety

At this stage, there is no certainty how to best reconcile the classical theory of relativity with quantum mechanics. Applying the well-tested methods of quantization to gravity – defined by the Einstein–Hilbert action – and quantizing the fluctuations around a classical solution to Einstein's equations leads to a nonrenormalizable theory. This happens because in four spacetime dimensions the mass dimension of the gravitational coupling constant G (in units where $\hbar$ and c are 1) is -2, whereas it should be larger than or equal to 0 for the theory to be renormalizable perturbatively. One would therefore expect the perturbative effective quantum field theory description to break down at energies E satisfying $GE^2 \gtrsim 1$.

There are of course well-known examples where the nonrenormalizability of a quantum field theory in the ultraviolet (UV) was eventually resolved by introducing new degrees of freedom, missed initially because they were not directly observable at low energies. The electroweak theory is an example where perturbative renormalizability was *regained* in this way. The theory was first described by a four-fermion interaction with an associated Fermi coupling G_F of mass dimension -2, just like the Newton constant G in gravity. As a result, its perturbation theory breaks down at energies with $G_F E^2 \gtrsim 1$. However, it turns out that for energies above $1/\sqrt{G_F} \approx M_W$, the mass of the W-particle, the four-fermion theory has to be replaced by the $SU(2)$-gauge theory of the weak interactions, which contains new excitations, the W- and Z-bosons. The new electroweak theory *is* a renormalizable quantum field theory.

Similarly, in the 1960s the low-energy scattering of pions was described by a nonlinear sigma model, another nonrenormalizable quantum field theory whose coupling constant, the pion decay constant F_π-squared, has mass dimension -2. However, high-energy scattering at energies beyond $1/F_\pi$ is no longer described well by the nonlinear sigma model, because it starts probing the intrinsic structure of the pions. A correct description has to incorporate appropriate new degrees of freedom, the quarks and gluons, and the corresponding quantum theory – quantum chromodynamics – is perfectly renormalizable.

There is no obvious reason which prevents us from writing down a perturbative (and nonrenormalizable)

expansion for gravity around some classical background geometry, say, flat Minkowski spacetime, if we are interested in an effective quantum field-theoretic description whose range of applicability does not extend beyond energies with $GE^2 \approx 1$. In view of the examples cited earlier, it is then tempting to conjecture that the apparent nonrenormalizability of gravity could be resolved by the appearance of new degrees of freedom at higher energies, rendering the theory renormalizable after all.

A solution of this kind may be in the form of a superstring theory in a higher dimensional spacetime, where the gravitational excitations are intertwined with infinitely many new degrees of freedom in such a way as to cure the UV problem. Although string theory cannot be ruled out as the correct answer, the world picture it provides has yet to be verified. In particular, supersymmetry – predicted by string theory – has not yet been observed at the Large Hadron Collider. Of course, even if no evidence of supersymmetry is found at this or future colliders, it may still be present at even higher energies. In this sense, the absence of observational evidence for supersymmetry does not disprove superstring theory as such, although it makes it less compelling as a resolution of the problem of unifying gravity and quantum theory.

There are other potential resolutions to the problem of finding a suitable *ultraviolet completion* of perturbative quantum gravity, which are not based on fundamental, string-like excitations and do not obviously require the existence of supersymmetry or extra dimensions. These are so-called nonperturbative approaches, whose starting point typically consists of a set of dynamical degrees of freedom closely modeled on those of classical gravity (*curved geometry* in one way or other), together with a nonperturbative prescription for quantization. A concrete example, that of *Causal Dynamical Triangulations*, will be described in some detail below. Its geometric degrees of freedom, in the presence of a UV cut-off, are given in terms of triangulated, piecewise flat spacetimes with discrete curvature assignments. Its nonperturbative quantization follows that of a standard lattice field theory, albeit with a dynamical rather than a fixed lattice.

An obvious charm of such a purely quantum field-theoretic ansatz lies in its minimalism, and the absence – to a large degree – of free parameters and other *tunable* ingredients. On the other hand, a key difficulty of this type of approach is to demonstrate that it is related to classical gravity in a suitable limit, something that is not at all obvious once one has moved beyond linearized quantum fields on a fixed background spacetime. One also needs to spell out what it means for the nonperturbative theory to *exist*, which likewise is nontrivial in a background-free description where *observables* are hard to come by.

In parallel with advances in string theory, also research in the wider area of nonperturbative quantum gravity has seen a steady rise in interest in recent decades. On the one hand, this was due to the rejuvenation of canonical quantum gravity in the form of *loop quantum gravity* from the late 1980s onward. (Curiously, this ansatz also postulates the fundamental character of certain one-dimensional *closed-string* (a.k.a. *loop*) excitations in the quantum theory.) At about the same time, the covariant *gravitational path integral* was given a new nonperturbative lease of life in terms of *dynamical triangulations*. Motivated originally by the search for a nonperturbative dynamics of curved, two-dimensional worldsheets in (bosonic) string theory, this dynamical lattice formulation provides a powerful computational tool for evaluating gravitational path integrals quantitatively: analytically in two, and numerically in higher dimensions. The focus of this chapter will be on this latter development, arguably the conceptually most straightforward and methodologically minimalist extension of the standard perturbative and covariant quantum field-theoretic formulation of gravity. We will explain how it may lead to the construction of a viable theory of quantum gravity, valid on all scales, without running into contradictions vis-à-vis the perturbative nonrenormalizability of the theory.

In the late 1970s, *Weinberg* outlined a scenario, coined *asymptotic safety* [34.1], for how quantum field theories which are not power-counting renormalizable around a trivial Gaussian fixed point could under certain, general conditions still make sense, just like ordinary renormalizable theories. In particular, an asymptotically safe theory is characterized by only a *finite* number of coupling constants, whose values will be determined by comparison with experiment or observation. The asymptotic freedom scenario is naturally described in the language of quantum field theory and the renormalization group. It is characterized by the presence of an ultraviolet fixed point in the infinite-dimensional coupling constant space of a theory, with the property that in the fixed point's neighborhood the dimension of the subspace of attraction is infinite-dimensional, with finite codimension. This codimension coincides with the number of free parameters of the theory that need to be fixed by experiment. Such

a UV fixed point therefore attains a similar status to that of the Gaussian fixed point of a renormalizable theory. The snag is that the tools of the perturbative theory are usually not sufficient to find such ultraviolet fixed points – if they exist for a given theory – and to study their neighborhoods.

To illustrate the implications of the presence of such a fixed point (in a somewhat simplistic fashion), let us introduce the dimensionless coupling

$$\tilde{G}(E) := GE^2. \quad (34.1)$$

A fixed point in this context always refers to the behavior under a change of scale E of a dimensionless, energy-dependent function like $\tilde{G}(E)$. The dimensionful quantity G in (34.1) can at this stage still be thought of as a (classical, low-energy) coupling *constant* of mass dimension −2. Let the behavior of $\tilde{G}(E)$ be dictated by a beta function $\beta(\tilde{G})$ according to

$$E\frac{\mathrm{d}\tilde{G}}{\mathrm{d}E} = \beta(\tilde{G}), \quad \text{with } \beta(\tilde{G}) = 2\tilde{G} - 2\omega\tilde{G}^2, \quad (34.2)$$

for some real parameter ω. It is immediately clear that for $\omega \neq 0$, $G = const$ is no longer a solution to (34.2). For consistency, G has to acquire a nontrivial E-dependence and therefore becomes a *function* $G(E) = \tilde{G}(E)/E^2$. In (34.2) we have chosen the simplest nontrivial beta function such that (i) in the limit of low energy, $E \to 0$, $G(E)$ goes to a constant (which we will continue to call G), and (ii) for $E \to \infty$, $\tilde{G}(E)$ goes to a nontrivial UV fixed point. Explicitly, the solution to the differential equation in (34.2) can be stated as

$$G(E) = \frac{G}{1 + \omega G E^2}, \quad (34.3)$$

from which we can read off the location of the UV fixed point at $\tilde{G} = 1/\omega$, the nontrivial zero of the beta function. An important feature of this solution is that the coupling constant $G(E)$ goes to zero at the UV fixed point.

In case the above example should appear somewhat ad hoc, it can be understood as arising from a more general construction, which starts from an asymptotically *free* theory in d dimensions. Figure 34.1a illustrates the corresponding (negative) beta function of the coupling g, together with a Gaussian UV fixed point at $g = 0$. If this theory is *lifted* to $d + \varepsilon$ dimensions – assuming that such a perturbation in the dimension is well defined, at least for small $\varepsilon > 0$ – its beta function will change according to

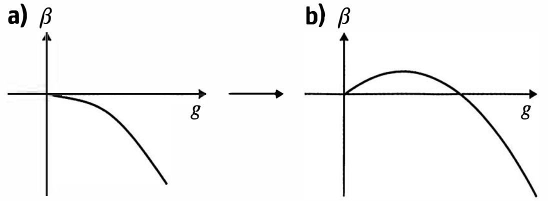


Fig. 34.1a,b Changing an asymptotically free theory to an asymptotically safe one by increasing its dimension from d to $d + \varepsilon$ results in a shift of its ultraviolet fixed point to a value $g > 0$

$$\beta(g) \to \rho(\varepsilon)g + \beta(g), \quad (34.4)$$

where $\rho(\varepsilon)$ is the (positive) amount by which the mass dimension of the coupling g *de*creases as a result of the dimensional increase by ε. (Our previous example, whose beta function was defined in relation (34.2), corresponds to $\rho = 2$.) Note that the Gaussian UV fixed point of the original theory has become a nontrivial UV fixed point away from zero in the higher dimensional theory, while $g = 0$ has been turned into an infrared fixed point, as illustrated by Fig. 34.1.

The theories we have discussed so far – four-Fermi theory, nonlinear sigma model and Einstein gravity – display a similar behavior in the sense that they are asymptotically free, renormalizable theories in spacetime dimension $d = 2$. Trying to make sense of them beyond dimension 2 by way of a $2+\varepsilon$-expansion, one encounters the situation depicted in Fig. 34.1. Of course, one may formally set $\varepsilon = 2$ in such an expansion, as would be needed to reach the dimension $d = 4$ of physical spacetime, but the validity of the perturbative expansion for such large values of ε would need to be established to take the results seriously, and a priori appears perhaps rather doubtful.

Nontrivial UV-complete extensions to $d = 4$ of the four-Fermi interaction or the nonlinear sigma model are not known and presumably do not exist. As mentioned above, we should rather think of them as effective theories, which happen to describe certain low-energy properties of more fundamental theories with more and different fundamental excitations. Still, it is difficult to draw any conclusions from this for general relativity, the theory we are interested in, which is after all very different physically: exactly the degrees of freedom that are fixed in all other theories, those of

spacetime itself, become dynamical in gravity. Much work has gone into trying to show that four-dimensional gravity possesses an ultraviolet fixed point with the requisite properties, either in terms of the $2+\varepsilon$-expansion [34.2–6] or by using general renormalization group techniques [34.7–12].

In what follows, we will not be concerned with the details of these efforts, but with the question of how the hypothesis of asymptotically safe gravity may be tested independently and nonperturbatively by using standard field-theoretic tools and by formulating quantum gravity via a lattice regularization.

34.2 A Lattice Theory for Gravity

A number of issues have to be addressed when representing gravity on a lattice. Is it possible in principle to construct a well-defined lattice regularization of gravity with a UV lattice cut-off, which can be removed in a controlled way to obtain a continuum limit (whatever this may turn out to be)? The answer is *yes*. More precisely, the issue is not so much how to represent gravity on a lattice, but how to represent a theory as a lattice theory whose standard continuum formulation in terms of local fields is diffeomorphism invariant, a vast gauge invariance closely related to the differentiable structure of the underlying manifold and its description in terms of local coordinate charts.

For the geometric degrees of freedom of the gravitational theory this can be done by viewing the lattice itself as representing directly a (piecewise linear) geometry. The key point is that such a geometry can be described uniquely without ever introducing coordinates, thus circumventing the associated redundancy of having to choose any particular set of coordinates. A convenient choice is to use lattices which are triangulations, in the sense of consisting of d-simplices, triangular building blocks which are d-dimensional generalizations of flat triangles ($=2$-simplices). Assuming the interior of a d-simplex to be flat, its geometry is uniquely specified by giving the lengths of its $d(d+1)/2$ one-dimensional edges or links. Together with the information of how the simplices are *glued together* (that is, how $(d-1)$-dimensional boundary simplices are identified pairwise) to form a triangulated manifold, this suffices to compute all geometric information, including distances, geodesics, volumes etc. without using coordinates. Important for our path integral representation, Regge observed that the curvature of such a piecewise linear geometry is in a natural way located on its $(d-2)$-dimensional subsimplices (the *hinges*). By the same token, the scalar curvature term of the Einstein action of such a geometry is given by the sum over all hinges of the deficit angle around each hinge, multiplied by the hinge's volume [34.13].

In our construction of a theory of quantum gravity, the lattice-regularized path integral over geometries thus becomes the sum over such triangulations, with weight depending on the Regge implementation of the Einstein action. Precisely which class of triangulations should we sum over in the path integral? When applying Regge calculus to classical gravity one uses a fixed lattice, in the sense of leaving the connectivity of its constituent simplicial building blocks unaltered. This still allows the curvature of the triangulation to be changed – for example, to optimally approximate that of a given smooth geometry – by changing the lengths of its one-dimensional edges.

When using the piecewise linear geometries in a path integral, the task is different. Firstly, we do not expect the individual path integral configurations to be smooth, but only continuous, in the same way as the paths in the path integral of a quantum-mechanical particle are continuous but in general nonsmooth (in fact, with unit probability they are nowhere differentiable). Similarly, the piecewise linear geometries are a subset of all continuous spacetime geometries. Note that we can even restrict ourselves to a subset of piecewise linear geometries as long as it is suitably dense in the set of all geometries. More precisely, when the lattice spacing goes to zero, we require the expectation values of observables, again suitably defined on the piecewise linear geometries, to converge to the value they would take in the continuum quantum field theory (which we assume exists). In contrast with the aim of the classical theory, we are therefore not trying to approximate any *particular* geometry by our lattice geometries, but to span the whole set of geometries.

In this context a specific subset of piecewise linear geometries has proved to be very useful, namely, the triangulations whose edges have all the same length a say. One can characterize this set of geometries as being constructed from gluing together *equilateral* simplicial building blocks in all possible ways, compatible with certain constraints (typically, a fixed topology and fixed

boundary components). Consequently, the variation in geometry (the way in which the geometric degrees of freedom are encoded) is linked to the mutual connectivity of the building blocks created by the gluing and not to variations in the link lengths, giving rise to the name *dynamical triangulations* (DT) [34.14–19]. From a path-integral perspective this approach has the advantage that distinct triangulations correspond to physically distinct geometries. Summing over this DT ensemble of geometries may therefore lead directly to the correct continuum measure in the limit that the UV cut-off is taken to zero, $a \to 0$. By contrast, treating the triangulations classically à la Regge, with fixed lattice connectivity and variable link lengths, still contains redundancies, in the sense that many different lattice configurations can correspond to the same physical geometry (see [34.20] and references therein). For illustration, consider a rectangle in the two-dimensional plane and triangulate its interior. Clearly, the interior vertices can be moved around locally in the plane without changing the flat geometry of the rectangle. However, since all of these are different as Regge triangulations, this leads to a severe overcounting in the path integral of quantum Regge calculus, for which there is currently no known fix.

Most importantly, the viability of the DT lattice regularization has already been demonstrated in a nontrivial case, that of gravity (coupled to matter) in two dimensions. As mentioned above, two-dimensional gravity is a renormalizable quantum field theory and various observables can be calculated analytically [34.21–23]. The dynamically triangulated two-dimensional lattice theory can also be solved, a number of observables can be calculated analytically and its continuum limit, taking the lattice spacing $a \to 0$, can be taken [34.24, 25]. Remarkably, results from the two different calculations can be compared and are found to agree. We conclude that *it is possible to provide a viable lattice regularization of a diffeomorphism-invariant quantum theory of geometries.*

One may object that this two-dimensional theory has little to do with true gravity in four spacetime dimensions; to start with, it has no propagating gravitons. However, we would like to argue that it is much more a theory of fluctuating geometries than one would ever expect of the four-dimensional theory. Because there is no Einstein–Hilbert action in two dimensions (it is topological), each configuration contributes in the path integral with the same weight, which is a maximally quantum situation. This is borne out by the analytic solutions of this model, which show the two-dimensional geometries as wildly quantum fluctuating. Nevertheless the lattice theory has no problem in reproducing the correct diffeomorphism-invariant continuum theory, also known as quantum Liouville gravity.

34.2.1 Observables

How to define what does and does not constitute an *observable* in quantum gravity, and how to construct and evaluate observables in any given formulation are physical questions of central importance. What we would like to highlight here is that a beautiful aspect of a geometric lattice formulation of quantum gravity of the type we are considering is that it forces one to address such questions head-on. It is not possible to hide behind some *expansion around flat spacetime*, but one is forced to think in terms of physical *rods and clocks*, much in the spirit of Einstein's classical theory.

Let us discuss the basic objects of any quantum field theory, namely, the correlators of local quantum operators $\mathcal{O}(x)$. Such correlators are important ingredients in constructing S-matrix elements, i. e., observables in quantum field theory on a fixed background. Also in conventional lattice theories, correlators play a crucial role in showing that a lattice theory has a continuum limit when the lattice spacing goes to zero.

Consider some lattice scalar field theory, and let $\mathcal{O}(x_n)$ be an operator at lattice spacetime coordinate $x_n = n \cdot a$, where a is the lattice spacing and n the integer-valued lattice coordinate. In general, we expect the correlator to fall off exponentially,

$$-\log\langle\mathcal{O}(x_n)\mathcal{O}(x_m)\rangle \approx \frac{|n-m|}{\xi(g_0)} + o(|n-m|), \quad (34.5)$$

where g_0 is the bare lattice coupling and $\xi(g_0)$ the correlation length in lattice spacings. The standard procedure for a lattice system is to take the continuum limit at a second-order phase transition point g_0^c, where the correlation length diverges like

$$\xi(g_0) \propto \frac{1}{|g_0 - g_0^c|^\nu}, \quad a(g_0) \propto |g_0 - g_0^c|^\nu. \quad (34.6)$$

Equation (34.6) tells us at what rate we should scale the lattice spacing to zero in the limit $g_0 \to g_0^c$, in order to find an exponential decay in the continuum, when the lattice correlation diverges, but the (dimensional) physical length $x_n - x_m = (n-m)a$ is kept constant,

$$m_{\text{ph}} a(g_0) = \frac{1}{\xi(g_0)},$$
$$\mathrm{e}^{-|n-m|/\xi(g_0)} = \mathrm{e}^{-m_{\text{ph}}|x_n - x_m|}. \quad (34.7)$$

Equation (34.7) illustrates the fact that dimensionful observables, like the physical mass m_{ph}, are defined by the *approach* to the critical point, not *at* the critical point.

The existence of a critical point and an associated divergent correlation length constitute the backbone of the Wilsonian renormalization group approach to quantum field theory. Since we are appealing to this Wilsonian approach by asking whether asymptotic safety is realized, it is important to understand whether it can be applied to quantum gravity at all. A first step in this direction is to understand whether suitable correlators and a correlation length can be defined in a diffeomorphism-invariant theory like quantum gravity. To start with, how can we define the distance between two points in a path integral where we integrate over the geometries defining this distance?

In flat d-dimensional spacetime, let us rewrite the correlator of a scalar field $\phi(x)$, say, in the form

$$\begin{aligned}&\langle\phi\phi(R)\rangle_V\\&\equiv\frac{1}{V}\frac{1}{s(R)}\int\mathcal{D}\phi\mathrm{e}^{-S[\phi]}\\&\times\int\mathrm{d}^dx\int\mathrm{d}^dy\phi(x)\phi(y)\delta(R-|x-y|).\end{aligned} \tag{34.8}$$

As indicated, this expression depends on a chosen distance R, but no longer on specific points x and y, which instead are integrated over. The integrand can be read *from right to left* as first averaging over all points y at a distance R from some fixed point x, normalized by the volume $s(R)$ of the spherical shell of radius R, and then averaging over all points x, normalized by the total volume V of spacetime. We assume translational and rotational invariance of the theory and that V is so large that we can ignore any boundary effects related to a finite volume.

This definition of a correlator is of course nonlocal, but unlike the underlying locally defined correlator has a straightforward diffeomorphism-invariant generalization to the case where gravity is dynamical, namely,

$$\begin{aligned}\langle\phi\phi(R)\rangle_V\equiv&\frac{1}{V}\int\mathcal{D}[g]\int\mathcal{D}_{[g]}\phi\mathrm{e}^{-S[g,\phi]}\delta\\&\times\left(V-\int\mathrm{d}^dx\sqrt{\det g}\right)\\&\times\int\mathrm{d}^dx\int\mathrm{d}^dy\frac{\sqrt{\det g(x)}\sqrt{\det g(y)}}{s_{[g]}(y,R)}\\&\times\phi(x)\phi(y)\delta(R-D_{[g]}(x,y)),\end{aligned} \tag{34.9}$$

which now includes a functional integration over geometries (in accordance with standard notation, $[g]$ denotes an equivalence class of metrics g under the action of the diffeomorphism group.) $[g]$, and dependences of the action, measures, distances, and volumes on $[g]$. Can the definition (34.9) be implemented meaningfully to define correlators in a quantum gravity theory? The answer is yes, and a two-dimensional example can again be used to demonstrate this. Namely, there are analytic predictions for the behavior of the propagators of certain matter theories coupled to two-dimensional Euclidean gravity [34.21–23], which have been shown to be reproduced by numerical simulations of the corresponding lattice theory [34.26–28]. By the way, their behavior is quite different from that of the flat space correlators, another manifestation of the fact that two-dimensional gravity is a theory of strong geometric fluctuations.

34.2.2 Time-Slicing and Baby Universes

An interesting aspect that can be analyzed in detail in the solvable two-dimensional quantum theory of fluctuating geometry is that of proper time. One usually considers a situation where the rotation to Euclidean signature has taken place and *proper time* is simply given by *geodesic distance*. In this setting, a closed one-dimensional spatial universe of fixed *time* is simply a loop of length ℓ. In the corresponding quantum theory one can ask for the amplitude for a universe of length ℓ_1 to *propagate* to another one of length ℓ_2 in proper time t. More precisely, the outgoing loop of length ℓ_2 is said to have a proper-time (in this case a geodesic) distance t to the incoming loop of length ℓ_1 if each point on ℓ_2 has geodesic distance t to ℓ_1. (The geodesic distance from a point to a set of points is defined as the minimum of the geodesic distances from the point to the points in the set.)

Figure 34.2 shows a typical geometry in the path integral contributing to the corresponding amplitude $G(\ell_1,\ell_2;t)$. It will often be convenient to work with its Laplace transform,

$$G(x,y;t)=\int_0^\infty\int_0^\infty\mathrm{d}\ell_1\,\mathrm{d}\ell_2\mathrm{e}^{-x\ell_1-y\ell_2}G(\ell_1,\ell_2;t). \tag{34.10}$$

We can view x and y in this expression as *boundary cosmological constants*, since $x\cdot\ell$ would be the action of

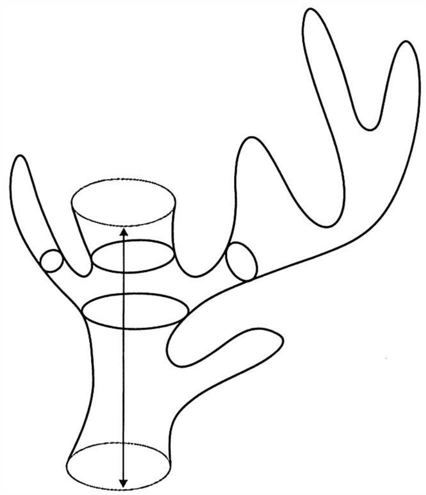


Fig. 34.2 Incoming and outgoing boundary loops of length ℓ_1 and ℓ_2, separated by a geodesic distance t, and a typical interpolating geometry of cylinder topology which contributes to the amplitude $G(\ell_1, \ell_2; t)$ in Euclidean signature. The additional loops drawn onto the interior geometry consist of points which share the same distance to the incoming loop. As indicated by the *upper set of three loops*, there can be many disconnected loops at a given distance to the incoming loop

a one-dimensional *spacetime* of volume ℓ and cosmological constant x.

As shown in [34.29], the amplitude $G(x, y; t)$ satisfies the remarkably simple equation

$$\frac{\partial G(x,y,t)}{\partial t} = \frac{\partial (W(x)G(x,y,t))}{\partial x}, \tag{34.11}$$

where $W(x)$ is the Hartle–Hawking disk amplitude, which in two-dimensional Euclidean gravity is given by [34.24, 25]

$$W(x) = \left(x - \frac{1}{2}\right)\sqrt{x + \sqrt{\Lambda}}. \tag{34.12}$$

As is clear from Fig. 34.2, space can branch out into many disconnected parts (i. e., change its topology) as a function of proper time t, giving rise to *baby universes*. The appearance of baby universes on all scales leads to the two-dimensional quantum spacetime being fractal, with Hausdorff dimension $d_h = 4$ [34.29, 30].

Rather amazingly, it is possible to integrate analytically over these baby universes, resulting (for each time history) in a spacetime with a proper-time foliation and no baby universes [34.31]. Alternatively, the expression for the loop–loop propagator without baby universes can be obtained directly by summing over a class of two-dimensional spacetimes which from the outset lack baby universes, provided one redefines the coupling constants suitably [34.32]. This latter procedure can also be implemented at the regularized level in terms of a set of *causal dynamical triangulations* (CDT), to be distinguished from the larger class of merely *dynamical triangulations* (DT), which served as carrier space for the Euclidean gravitational path integral [34.32].

The resulting theory has a well-defined Hamiltonian and corresponding unitary proper-time evolution. The explicit map between the cosmological constants of DT and CDT turns out to be nonanalytic,

$$\tilde{\Lambda}_{\mathrm{CDT}} = \sqrt{\Lambda_{\mathrm{DT}}}\,, \quad \tilde{x}_{\mathrm{CDT}} = \sqrt{x + \sqrt{\Lambda_{\mathrm{DT}}}}, \tag{34.13}$$

where we have denoted the CDT-analogs of the couplings with a subscript and tilde. Consequently, in CDT both lengths and areas acquire a dimensionality different from that found in the DT ensemble of spacetimes and in Liouville gravity. When using the CDT ensemble, also the Hausdorff dimension changes from 4 to 2, the canonical value for ordinary smooth two-dimensional spacetimes. (A word of warning: the coincidence in Hausdorff dimension does *not* allow one to conclude that the quantum geometry of two-dimensional CDT in any way approximates a smooth classical manifold; in fact, it does not.)

The CDT loop–loop propagator satisfies the equation

$$\frac{\partial \tilde{G}(\tilde{x},\tilde{y},t)}{\partial t} = \frac{\partial ((\tilde{x}^2 - \tilde{\Lambda}_{\mathrm{CDT}})\tilde{G}(\tilde{x},\tilde{y},t))}{\partial \tilde{x}}, \tag{34.14}$$

and the Hamiltonian governing the (proper-) time evolution is given by

$$\tilde{G}(\tilde{\ell}_1, \tilde{\ell}_2, t) = \langle \tilde{\ell}_2 | \mathrm{e}^{-t\hat{H}} | \tilde{\ell}_1 \rangle\,,$$
$$\hat{H} = -\tilde{\ell}\frac{\mathrm{d}^2}{\mathrm{d}\tilde{\ell}^2} + \tilde{\Lambda}_{\mathrm{CDT}}\tilde{\ell}, \tag{34.15}$$

while the CDT Hartle–Hawking wave function (which is derived from the propagator $\tilde{G}$ [34.32]) satisfies

$$\hat{H}\tilde{W}_{\mathrm{CDT}}(\tilde{\ell}) = 0\,. \qquad \text{(Wheeler–DeWitt)} \tag{34.16}$$

Above, our first way of deriving this formulation was as a kind of *effective* theory: we started from

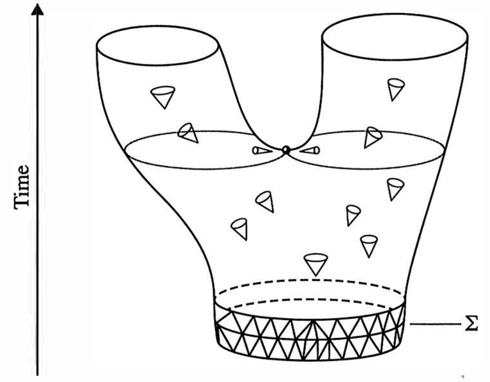


Fig. 34.3 The light cone structure (and therefore the underlying Lorentzian geometry) becomes degenerate in points where space splits in two

the set of all Euclidean two-dimensional geometries of a fixed topology. These geometries are *isotropic* in the sense that they do no carry any a priori preferred direction. We then superimposed a notion of proper time on them and integrated out part of the degrees of freedom. However, when starting in the physically correct *Lorentzian* signature, one can formulate a general principle which excludes geometries whose *spatial* topology is not constant in time [34.33, 34]. The point is that spatial topology changes are associated with causality violations of one kind or other. This is illustrated by the *trouser geometry* depicted in Fig. 34.3. As is clear from the embedding of this two-dimensional spacetime in flat Minkowski space, with time pointing upward, there must be at least one point near the crotch of the trousers where the tangent plane is exactly horizontal and the light cone therefore degenerate. Note that imposing causality conditions on the geometry to eliminate such configurations only makes sense in the presence of a Lorentzian metric and cannot even be formulated in a purely Euclidean theory, in the absence of any extra structure.

By the same token, one can take as domain of the path integral the set of all Lorentzian piecewise flat triangulations whose causal structure is well defined, and where in particular no changes of spatial topology are allowed to occur. The set of CDT – which can be defined in any dimension (not just $d=2$) – obeys a strong version of causality of this kind, which is implemented by requiring each triangulation to be the product of a one-dimensional *triangulation* (a line with equidistant points), representing discrete proper time, and other triangulated degrees of freedom, representing the spatial directions of the geometry, which may be thought of as triangulated fibres over a one-dimensional base space. (Product triangulations, of which this is a particular instance, were investigated in [34.35], see also [34.36].) As an added bonus, each triangulation in the class of CDT can be analytically continued to Euclidean signature, and the associated gravitational Regge actions satisfy the standard relation between actions defined in spacetimes of Lorentzian and Euclidean signatures, namely,

$$\mathrm{i}S_{\text{Lorentzian}} \mapsto -S_{\text{Euclidean}}. \tag{34.17}$$

Despite the fact that the actions obey (34.17), the Lorentzian theory defined on CDT geometries will even after this *Wick rotation* be *distinct* from the full Euclidean theory, because not every Euclidean triangulation is the image of a causal, Lorentzian one. The subclass of Euclidean geometries that are in the image can be obtained *surgically* as explained above, by superimposing a notion of proper time on each Euclidean triangulation and then removing all of its baby universes associated with spatial topology changes. The two-dimensional case is sufficiently simple to allow us to perform the calculation in either way, by starting from a path integral over all Euclidean geometries and removing baby universes, or by starting from a path integral over causal (CDT) geometries and rotating it to Euclidean signature. Both results agree after a redefinition of the coupling constants. Let us note in passing that our formulation – not only in dimension 2, but also in higher dimensions – has a couple of characteristics reminiscent of so-called Hořava–Lifshitz gravity, namely, the use of a preferred time foliation and a unitary time evolution. We will return to this subject in Sect. 34.4 below.

34.2.3 CDT in Higher Dimensions

It is not known whether the above-described procedure of integrating out baby universes in $d=2$ can be generalized to higher dimensions in a simple and useful way. It implies that at this stage, we have two a priori unrelated lattice gravity theories in dimension $d>2$, one purely Euclidean based on DT and one Lorentzian based on CDT. The latter starts out in physical, Lorentzian signature, and imposes local causality conditions (nondegeneracy of local light cones) and a proper-time time foliation. (Note that there is no strict physical requirement that individual path integral histories *must* be causal; individual histories are not

physical, observable quantities, only expectation values computed in the ensemble of histories are.) For calculational purposes, these lattice configurations are then rotated to Euclidean signature and the path integral over this class can in principle be performed. Of course, since the physics one hopes to describe ultimately by these theories has Lorentzian character, one will have to perform an *inverse Wick rotation* back to Lorentzian spacetime eventually, never mind whether the computation at an intermediate step took place in a purely Euclidean or in a Euclideanized Lorentzian framework.

The simplest implementation of Euclidean DT based on the lattice Regge version of the Einstein–Hilbert action (the inclusion of a cosmological term being understood) does not seem to lead to a theory with an interesting continuum limit. Even if this is the case, it is in principle possible that by adding more terms to the bare lattice action and suitably tuning the associated new coupling constants, an interesting continuum theory may emerge after all. This possibility has been investigated in the past [34.37–39], as well as more recently [34.40, 41], but there is no conclusive evidence at this point that these modified Euclidean models can reproduce the physical properties of quantum gravity from CDT, the Lorentzian lattice gravity theory to which we will turn next (see also [34.42–50] for a variety of reviews of the subject).

Figure 34.4 illustrates the general construction of a four-dimensional CDT triangulation. We take space to be compact and with the simplest topology, that of the three-sphere S^3. In addition, we assume a discrete proper-time foliation and represent the spatial geometry at each integer proper time t by a three-dimensional simplicial manifold, given as some configuration of Euclidean DT in terms of equilateral tetrahedra. By assumption, the tetrahedra are flat in the interior, which means that their geometric properties are uniquely specified by the length of their edges, which is some number $a_s > 0$ (the same for all edges). To obtain a four-dimensional Lorentzian simplicial manifold with signature $(-+++)$, we must still fill in all intervals $[t, t+1]$ between consecutive spatial slices. This can be done by using two types of geometrically distinct four-simplices, which again by assumption are flat in the interior, but this time with Lorentzian signature. The two different types are the (4,1)- and the (3,2)-simplex depicted in Fig. 34.4, together with their time-reversed counterparts. The (4,1)-simplex has as its *base* one of the spatial tetrahedra contained in the triangulated constant-time slice. (The *4* in the label (4,1) refers to the four vertices contained in slice t that span this tetrahedron; similarly, the *1* refers to the single vertex shared with slice $t+1$. An analogous labeling has been used for the (3,2)-simplex.) All that remains to be done to fix the geometry of the four-simplices is to assign lengths to the edges that have their end points in adjacent slices, and whose time labels therefore differ by one unit. We choose them to be all time-like and of equal (absolute) length $a_t > 0$, which in our signature convention implies that their squared edge length is given by $-a_t^2$. (Note that a_t gives us an approximate distance measure between adjacent spatial slices labeled by integer-t, where the distance of a point in slice $t+1$ to slice t is defined as the length of the longest geodesic from the point to the slice.)

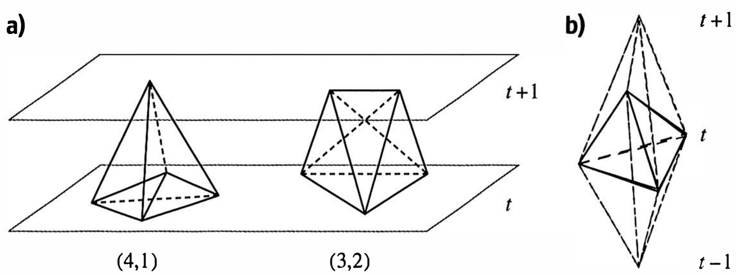


Fig. 34.4a,b A triangulation in CDT consists of four-dimensional triangulated layers assembled from (4,1)- and (3,2)-simplices, interpolating between adjacent integer constant-time slices (**a**), which in turn are triangulations of S^3 in terms of equilateral tetrahedra. Each purely spatial tetrahedron at time t forms the interface between two (4,1)-simplices, one in the interval $[t-1, 1]$, and the other in $[t, t+1]$, as illustrated on (**b**). Although a (3,2)-simplex shares none of the five tetrahedra on its surface with a constant-time slice (the tetrahedra are all Lorentzian), it is nevertheless needed in addition to the (4,1)-building block to obtain simplicial manifolds with a well-defined causal structure

Our choice of causal geometries and length assignments has the added benefit that we can define a map that uniquely maps each Lorentzian CDT history to a Euclidean DT history. Let us start by parameterizing the relative length of the two lattice parameters a_s and a_t by a positive real number α defined by $\alpha := -a_t^2/a_s^2$. Performing a rotation $\alpha \to -\alpha$ in the complex lower half plane can be interpreted as changing all time-like length assignments of lattice links to space-like ones according to

$$a_t^2 = -\alpha a_s^2 \;\to\; a_t^2 = \alpha a_s^2. \tag{34.18}$$

In order that the Euclidean four-simplices obtained after this rotation satisfy triangle inequalities we require $\alpha > 7/12$. The resulting triangulation represents a piecewise linear manifold with *Euclidean* signature. If one writes the Lorentzian Regge action as a function of a single lattice parameter $a := a_s$ and of α, the action behaves under the rotation (34.18) as one would expect naively from a rotation from Lorentzian to Euclidean spacetime, namely,

$$\mathrm{i}S_{\mathrm{L}}[\alpha] = -S_{\mathrm{E}}[-\alpha]. \tag{34.19}$$

The prescription (34.18) leading to (34.19) is the *Wick rotation* we had in mind in our earlier discussion in Sect. 34.2.2. It transforms the original Lorentzian path integral with complex weights $\mathrm{e}^{\mathrm{i}S_{\mathrm{L}}(T)}$ to one with real weights $\mathrm{e}^{-S_{\mathrm{E}}(T)}$, where by slight abuse of notation we use the same symbol T to denote the initial triangulation (with Lorentzian edge length assignments) and the one after rotation (which has identical connectivity, but purely Euclidean edge length assignments). Modulo the sign flip for the length assignments, the domain of the Euclideanized path integral is the same set $\mathcal{T} = \{T\}$ of triangulations as that of the original Lorentzian path integral. The set $\mathcal{T}$ is of course smaller than the set of *all* Euclidean triangulations one would obtain by gluing together the same Euclideanized building blocks, because it still carries an imprint of the causality conditions imposed on the Lorentzian triangulations.

The fact that in DT and CDT we use *standardized* building blocks to construct the triangulations means that the Regge action takes on a very simple functional form. For the special case $|\alpha| = 1$ we have after the Wick rotation only a single type of building block, the equilateral four-simplex with all link lengths equal to $a \equiv a_s$. The Regge form of the Einstein–Hilbert action becomes

$$S_{\mathrm{E}}[-\alpha = -1; T] = -\kappa_0 N_0(T) + \kappa_4 N_4(T), \tag{34.20}$$

as is well known from Euclidean DT quantum gravity. In (34.20), $N_0(T)$ denotes the number of vertices in the triangulation T, and $N_4(T)$ the number of its four-simplices. The coupling κ_0 is related to the gravitational coupling constant G via $1/\kappa_0 \propto Ga^2$, and κ_4 should be identified with $a^4\Lambda/G$, where Λ is the cosmological constant.

Whenever $|\alpha| \neq 1$, we retain the two different building blocks (of type (4,1) and (3,2)) after the rotation, and the action will depend on their total numbers, $N_4^{(4,1)}$ and $N_4^{(3,2)}$, separately instead of only on their sum $N_4 = N_4^{(4,1)} + N_4^{(3,2)}$. It is convenient to parameterize the resulting Euclideanized Regge action in the form

$$\begin{aligned} S_{\mathrm{E}}[-\alpha; T] = &-(\kappa_0 + 6\Delta)N_0(T) \\ &+ \kappa_4\left(N_4^{(3,2)}(T) + N_4^{(4,1)}(T)\right) \\ &+ \Delta\left(N_4^{(3,2)}(T) + 2N_4^{(4,1)}(T)\right), \end{aligned} \tag{34.21}$$

where the asymmetry parameter Δ is a function of α such that $\Delta(\alpha = 1) = 0$.

We note that Δ appears in (34.21) on a par with the other two coupling constants, κ_0 and κ_4. In what follows, we will treat it as a third independent coupling constant. The reason for doing this – despite the fact that it has no immediate interpretation in the Einstein–Hilbert action – is that in the region of phase space (the space spanned by the three couplings κ_0, κ_4 and Δ) where we observe interesting, apparently continuum physics, the entropy of geometries is as important as the contributions coming from the bare action term. To make this more explicit, one can rewrite the Euclidean partition function of the theory as a sum over the counting variables $N_4^{(4,1)}$, $N_4^{(3,2)}$ and N_0 according to

$$\begin{aligned} Z(\kappa_0, \kappa_4, \Delta) &= \sum_T \mathrm{e}^{-S_{\mathrm{E}}[T]} \\ &= \sum_{N_4^{(4,1)}, N_4^{(3,2)}, N_0} \mathrm{e}^{-S_{\mathrm{E}}\left[N_4^{(4,1)}, N_4^{(3,2)}, N_0\right]} \\ &\qquad \mathcal{N}\left(N_4^{(4,1)}, N_4^{(3,2)}, N_0\right), \end{aligned} \tag{34.22}$$

where $\mathcal{N}(N_4^{(4,1)}, N_4^{(3,2)}, N_0)$ is the number of triangulations with $N_4^{(4,1)}$ four-simplices of type (4,1), $N_4^{(3,2)}$ four-simplices of type (3,2) and N_0 vertices. Introducing the notation $c_1 = N_0/N_4^{(4,1)}$ and $c_2 = N_4^{(3,2)}/N_4^{(4,1)}$, the leading-order behavior of this combinatorial quantity in the large-volume limit is known to be of the

form

$$\mathcal{N}(N_4^{(4,1)}, N_4^{(3,2)}, N_0) = e^{f(c_1,c_2)N_4^{(4,1)} + \text{s.l.}}, \quad (34.23)$$

where *s.l.* denotes subleading terms in $N_4^{(4,1)}$, and the c_i typically have some boundedness properties. Since in the same limit the action (34.21) can be similarly approximated by $S_E = \tilde{f}(c_1, c_2)N_4^{(4,1)} + \text{s.l.}$, it implies that in the region of phase space where the four-volume can become large, both $\mathcal{N}$ and e^{-S_E} have the same functional form and are potentially of the same magnitude. It turns out that this is the same region where we observe interesting continuum-like physics. Because of contributions from both *energy* and *entropy*, it is clear therefore that the effective action governing physics in this nonperturbative region can be very different from the *naïve* Einstein–Hilbert action, justifying our inclusion of Δ as a tunable parameter in the bare action.

To summarize: taking as our starting point spacetimes with Lorentzian signature, we can consider the transition amplitude between an initial and a final spatial three-geometry, $[g_i^{(3)}]$ and $[g_f^{(3)}]$ separated by a proper time t. We can then regularize the theory, using CDT, representing three-geometries by equilateral Euclidean triangulations and spacetime geometries by causal, Lorentzian triangulations with a discrete proper-time foliation. In the CDT framework, each of the latter can be rotated to Euclidean signature, leading to a regularized, Euclideanized sum-over-histories. What remains to be done is to *remove the regulator*, that is, take the lattice spacing a to zero. Denoting the initial and final spatial triangulations by $T_i^{(3)}$ and $T_f^{(3)}$, we thus arrive at the prescription

$$\begin{aligned} &G_E\left(\left[g_i^{(3)}\right], \left[g_i^{(3)}\right], t, \kappa_0, \kappa_4, \Delta\right) \\ &:= \lim_{a\to 0} \sum_{T:T_i^{(3)}\to T_f^{(3)}} e^{-S_E[T]}, \end{aligned} \quad (34.24)$$

which can be viewed as the four-dimensional generalization of the two-dimensional loop–loop amplitude $\tilde{G}(\tilde{\ell}_1, \tilde{\ell}_2, t)$ introduced in (34.15). For a more detailed description of the CDT construction we refer the interested reader to [34.51–55].

34.3 The Phase Diagram

Contrary to the situation in two dimensions, we cannot calculate the amplitude (34.24) analytically. However, we can extract a lot of nontrivial, nonperturbative information by performing Monte Carlo computer simulations. This will usually start with an investigation of the structure of the space of coupling constants (the *phase space* of the underlying statistical system), in particular, trying to identify regions associated with a second-order phase transition, where according to standard lore one can hope to obtain continuum physics.

Let us highlight two technical aspects related to our implementation of the computer simulations. Firstly, rather than fixing specific boundary three-geometries $T^{(3)}$ at times 0 and t, we take time to be periodic. Although this is strictly speaking in contradiction with imposing causality (it introduces closed time-like curves), in practice it turns out to not affect results. The nature of the ground states of geometry is such that by choosing t sufficiently large – assumed from now on – the boundary condition becomes irrelevant.

Secondly, as we have discussed, action (34.21) depends on three coupling constants, one of which, κ_4, can be identified with the cosmological coupling constant, multiplying the spacetime volume V in the action. In the computer simulations it is convenient to keep this four-volume fixed, which means that the cosmological constant does not really play a role. We compensate for this by performing separate simulations at different (fixed) spacetime volumes. From these we can in principle reconstruct results which depend on the cosmological constant via a Laplace transformation,

$$G(\kappa_4, \ldots) = \int_0^\infty dV e^{\kappa_4 V} G(V, \ldots). \quad (34.25)$$

We are therefore left with two coupling constants, κ_0 and Δ. The corresponding phase diagram is shown in Fig. 34.5 [34.56] and exhibits three distinct phases, labeled A, B, and C. Phase C appears to be the one relevant for continuum physics, because only there do we observe extended four-dimensional universes [34.57, 58]. A careful numerical analysis reveals strong evidence that the transition between phases C and A is first order, whereas between phases C and B we find a second-order transition [34.59, 60]. This very exciting result implies that the B–C phase transition line is a candidate for a region in the coupling-constant plane

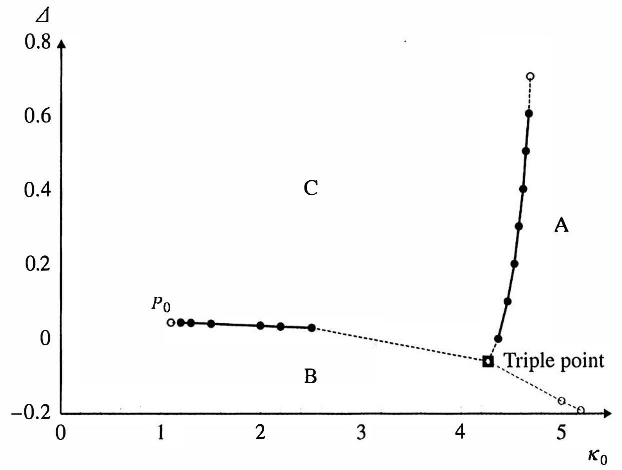


Fig. 34.5 The phase diagram of CDT quantum gravity in the (κ_0, Δ)-plane

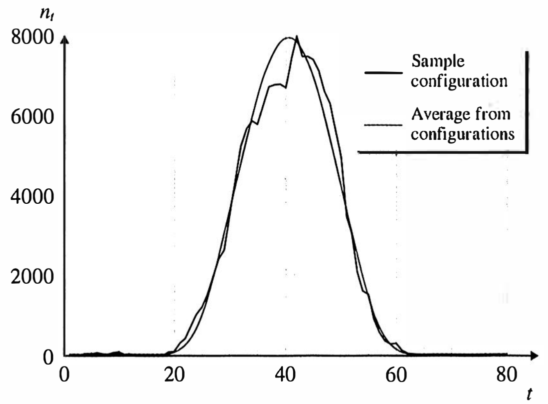


Fig. 34.6 The three-volume of spatial slices as a function of proper time in phase C. Shown are a sample configuration of the volume profile, as well as the expectation value of the same quantity

where genuine UV continuum limits may exist, defined by approaching specific points on the line. Conversely, moving away from the transition line into phase C corresponds to going toward an IR limit.

34.3.1 Phase C

The reason why phase C is related to extended four-dimensional spacetimes is illustrated in Fig. 34.6, which shows both a sample path-integral configuration generated by the computer during the Monte Carlo simulations, as well as the associated quantum observable, obtained by averaging in the ensemble. While of course we have access to the complete geometric information of the quantum spacetimes that are generated, only a single degree of freedom is depicted here, the three-volume of a spatial slice of the quantum spacetime as a function of proper time. The time extension in a given simulation is always fixed (in the case at hand to 80 discrete time steps). What we observe in Fig. 34.6 is that the universe does not make use of the full time interval available, but has a nonvanishing volume only on a connected subset of the time axis. (Since we impose the kinematical constraint that the spatial volume at fixed t cannot become smaller than 5 tetrahedra – the minimal number required to build a simplicial manifold of topology S^3 – the volume never vanishes completely. More precisely, what we observe in addition to the bell-shaped part of the volume profile is the formation of a distinct *stalk* which is close to the minimal size of 5 everywhere.)

A quantitative piece of evidence in favor of a *four-dimensional* extended universe is the fact that its time extension (not counting the stalk) scales like $N_4^{1/4}$ when the total discrete four-volume N_4 of the universe used in the simulations is varied. Similarly, its discrete three-volume $N_3(t)$ scales like $N_4^{3/4}$. Contrary to one's naïve expectations, these findings are highly nontrivial, because they have been derived in a nonperturbative, background-independent path integral formulation. The simplicial building blocks of our regularization *are* four-dimensional, but since assembling them is only dictated by the Boltzmann weight $e^{-S_E[T]}$ *without* any reference to a four-dimensional background, there is no reason why the resulting object, extrapolated to infinite lattice volume, should be four-dimensional on any scale.

This is specifically true in the nonperturbative regions of phase space where the entropic contributions to the effective action compete with those coming from the classical bare action, as explained above. In these regions it can easily happen that a type of configuration is entropically favored that has no resemblance at all with an extended four-dimensional universe. Just from looking at the volume profiles, it is obvious that something like this does indeed happen in phases A and B, which as a result do not appear to have any classical limit resembling general relativity [34.57, 58]. However, even in phase C the observed quantum universe is truly an outcome of nonperturbative dynamics, not a consequence of the dominance of the classical action. (Since we are working in Euclidean signature, domi-

nance of the classical action would be fatal for the path integral, because of the action's unboundedness from below. In phase C, this instability is cured by the entropy of *microstates* or, in other words, the path-integral measure [34.61–63].)

The fact that the path-integral measure can play a crucial role in determining the nonperturbative dynamics was a main lesson learned already earlier in the context of four-dimensional DT quantum gravity. When one considers a path integral ensemble of geometries obtained from gluing four-dimensional equilateral Euclidean simplices, with the only constraint that the topology should be that of S^4, one ends up with a universe of vanishing linear extension and infinite Hausdorff dimension [34.64]. This makes the situation depicted in Fig. 34.6 all the more remarkable!

34.3.2 The Effective Action

However, the surprises do not stop here. The smooth curve in Fig. 34.6 represents the expectation value of the volume profile, that is, the average over path integral configurations measured in the Monte Carlo simulations. For N_4 sufficiently large this curve is very precisely fitted by the function

$$\langle N_3(i)\rangle \propto N_4^{3/4}\cos^3\left(\frac{i}{s_0 N_4^{1/4}}\right), \tag{34.26}$$

where i denotes (integer) lattice time, N_4 the total number of four-simplices, and $N_3(i)$ the number of tetrahedra at time i [34.65, 66], and s_0 is a constant. (The formula is of course not valid in the stalk, where $N_3(i) \approx 5$.)

Can the functional form of the expectation value found in (34.26) be obtained directly from an action principle? The answer is yes [34.61]. A long time ago, Hartle and Hawking explored a minisuperspace approach to quantum gravity, where all gravitational (field) degrees of freedom at a fixed time are represented by a single number, the so-called scale factor or, equivalently, the total three-volume of the universe. (This rather crude approximation is borrowed from standard cosmology, where homogeneity and isotropy are assumed to give a realistic description of our universe on the very largest scales.) Taking this classically reduced formulation as the starting point of the quantization, finding a quantum theory of gravity is reduced to a quantum mechanical problem in one variable, the scale factor $a(t)$ [34.67].

The volume profile (34.26) of the emergent extended universe found in phase C of CDT quantum gravity can be derived from an *effective* action for the three volume, namely,

$$S_{\text{eff}} = \frac{1}{24\pi G}\int \mathrm{d}t\left(\frac{\dot{V_3}^2(t)}{V_3(t)} + k_2 V_3^{1/3}(t) - \lambda V_3(t)\right), \tag{34.27}$$

where t denotes the proper time, k_2 is the numerical constant, and λ is the Lagrange multiplier, not a cosmological constant, because the total four-volume V_4 is kept fixed in the simulations. Intriguingly, one obtains exactly the same expression (up to an overall sign) when plugging a spatially homogeneous and isotropic ansatz for the metric $g_{\mu\nu}(x)$ into the Euclidean Einstein–Hilbert action, and re-expressing the dependence on the scale factor in terms of the three-volume $V_3(t) \propto a^3(t)$. The solution to the equations of motion derived from (34.27) is the Euclidean de Sitter universe (a round four-sphere), which as a function of proper time t results in the $\cos^3(t/V_4^{1/4})$-dependence of (34.26).

Despite the fact that they lead to very similar results for the dynamics of the scale factor, let us stress that conceptually there is a big difference between the ansatz of Hartle and Hawking, who simply assumed a minisuperspace reduction from the outset, and studying the effective dynamics of (the expectation value of) the scale factor in a full theory of quantum gravity, as we are doing. The only small but important reminder of the nonperturbative origin of the action (34.27) is its overall sign, which is opposite to that found in Euclidean cosmology. It can be attributed directly to *entropic* contributions to the effective action. The solutions to the equations of motion are of course not affected by this sign difference. A discretization of the effective action (34.27) has the functional form

$$S_{\text{discr}} = k_1\sum_i\left(\frac{(N_3(i+1)-N_3(i))^2}{N_3(i)} + \tilde{k}_2 N_3^{1/3}(i) - \tilde{\lambda} N_3(i)\right). \tag{34.28}$$

We have managed to reconstruct it in detail from the simulation data for the volume–volume correlator $\langle V_3(t)V_3(t')\rangle$, and have also shown that the quantum fluctuations around the de Sitter *background geometry* are well described by the action (34.28), yet another nontrivial result [34.66].

The same data have allowed us to relate the continuum coupling constant G in (34.27) to the constant k_1 in (34.28) according to

$$G = \frac{a^2}{k_1} \frac{\sqrt{C_4} s_0^2}{3\sqrt{6}}, \tag{34.29}$$

where a is the lattice spacing and C_4 is essentially the volume of a four-simplex (for lattice spacing $a = 1$), but depends weakly on the ratio between $N_4^{(1,4)}$ and $N_4^{(2,3)}$ (since the (4,1)- and (3,2)-simplices only have identical four-volumes when $\alpha = 1$). This ratio, as well as the value of the constant s_0, defined in (34.26), depend on the choice of the bare coupling constants κ_0 and Δ in phase C.

Let us consider a typical choice for these couplings, $(\kappa_0, \Delta) = (2.2, 0.6)$, positioning us in the interior of phase C. At this point in phase space, we have measured k_1 and with the help of (34.29) expressed Newton's constant and the Planck length ℓ_P in terms of the lattice spacing, resulting in

$$G \approx 0.23a^2 , \quad \ell_P \equiv \sqrt{G} \approx 0.48a. \tag{34.30}$$

From the identification of spacetime with a Euclidean de Sitter universe we have that $V_4 = 8\pi^2 R^4/3 = C_4 N_4 a^4$, where C_4 is the same quantity that appeared in (34.29). For the range of four-volumes used in the simulations, $N_4 \in [45\,000, 360\,000]$, the linear size πR of the quantum de Sitter universes lies between 12 and 21 Planck lengths ℓ_P. The small size of our universes is compatible with the fact that the observed quantum fluctuations in the three-volume are quite substantial, as illustrated in Fig. 34.6 (see also Fig. 34.7). For larger universes, the volume fluctuations will quickly become irrelevant.

However, in order to investigate quantum properties of spacetime at Planckian and even sub-Planckian length scales, we want to do the opposite, namely, make the universes smaller and in this way increase the small-scale resolution of the simulations. How can we improve on (34.30) such that a single Planck length ℓ_P corresponds not to just half a lattice spacing, but to many lattice spacings a? From (34.29) and (34.30) it is clear that when k_1 goes to zero, ℓ_P can become much larger than a. The question is whether we can adjust k_1 to go to zero. Since k_1 depends on the bare coupling constants κ_0 and Δ, we have performed a scan of phase C to determine its qualitative behavior [34.66]. Moving toward the A–C phase transition, k_1 is indeed decreasing, without going all the way to zero in the range of coupling constants scanned so far. Approaching the B–C phase transition is more difficult, because the system undergoes a second-order transition, and we observe a corresponding critical slowing-down. As far as we can tell from the numerical data at this stage, k_1 does *not* decrease when we approach this transition. However, as we will see in the next section, having k_1 go to a fixed value different from zero is actually the behavior predicted at an ultraviolet second-order transition line, and therefore compatible with the continuum scenario we have appealed to earlier.

34.3.3 Making Contact with Asymptotic Safety

Let us return to the renormalization group (34.2), which was formulated in terms of the dimensionless coupling constant $\tilde{G} = GE^2$. Now that we have a UV cut-off, the lattice link length a, we can instead form the dimensionless quantity $\hat{G} = G/a^2$. From (34.29) it can essentially be identified with the inverse of k_1, which we can measure. We can reformulate the renormalization group in terms of the new short-distance cut-off as

$$\begin{aligned} G(a) &= a^2 \hat{G}(a) , \\ a\frac{\mathrm{d}\hat{G}}{\mathrm{d}a} &= -\beta(\hat{G}) , \\ \beta(\hat{G}) &= 2\hat{G} - c\hat{G}^2 + \cdots , \end{aligned} \tag{34.31}$$

where c depends on the constant ω of (34.2). Near the putative non-Gaussian UV fixed point $\hat{G}^*$, we can expand $\hat{G}$ and k_1 to the lowest order in a according to

$$\hat{G}(a) = \hat{G}^* - K a^{\tilde{c}}, \quad k_1(a) = k_1^* + \tilde{K} a^{\tilde{c}}, \tag{34.32}$$

for some $K, \tilde{K}$, where the approach to the fixed point is governed by the exponent

$$\tilde{c} = -\beta'(\hat{G}^*). \tag{34.33}$$

As explained in Sect. 34.2.1, in standard lattice theory one would now relate the lattice spacing near the fixed point to the bare coupling constants with the help of some correlation length ξ. However, in four-dimensional quantum gravity we do not yet have a suitable correlation length at our disposal which could play this role.

In search of an alternative, let us first consider the equation $V_4 = N_4 a^4$, which defines the dimensionful continuum four-volume V_4 in terms of the number N_4 of

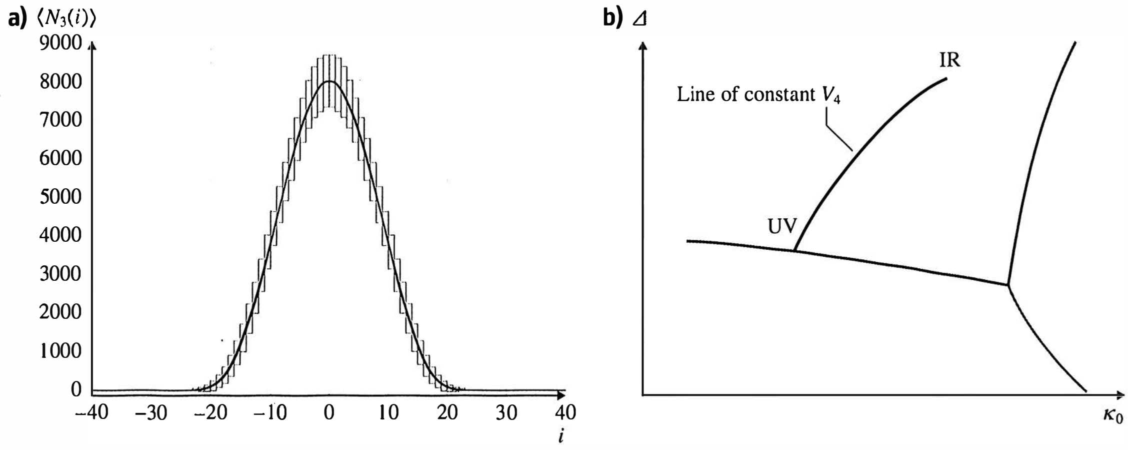


Fig. 34.7 **(a)** Three-volume profile for given N_4, for specific values (κ_0, Δ) of the bare coupling constants. Also indicated is the magnitude of the three-volume fluctuations around the mean value. While the expectation value of the three-volume scales like $N_4^{3/4}$, the fluctuations only scale like $N_4^{1/2}$. **(b)** Identifying a path of *constant physics* in the κ_0-Δ plane. Starting at some point in phase C, a path moving toward the UV phase transition is created by increasing N_4 and simultaneously adjusting κ_0 and Δ, such that the ratio of the size of the three-volume fluctuations and the expectation value of the three-volume remains constant

four-simplices and the lattice spacing. If we could consider V_4 as fixed, we could replace the a-dependence of (34.32) by a N_4-dependence, with the advantage that N_4 is a parameter we can straightforwardly control. Re-expressing (34.32) in terms of N_4 yields

$$k_{\mathrm{I}}(N_4) = k_{\mathrm{I}}^* - K' N_4^{-\tilde{c}/4}\,, \tag{34.34}$$

for some K'. Since we can measure k_{I}, we could determine the flow to the fixed point. The question is now which lattice measurements we should perform in order to make (34.34) applicable. Increasing N_4 while *staying* at a specific point (κ_0, Δ) in phase C does *not* correspond to keeping V_4 fixed, because during this process the size of the quantum fluctuations in the three-volume decreases relative to the expectation value of the three-volume. (More precisely, we already know that the ratio goes to zero like $1/N_4^{1/4}$.) Conversely, if *physics* is to be constant, which includes a constant V_4, that same ratio should also remain constant.

We will use this observation as our definition for what we mean by a *path of constant physics*. If we had a correlation length available, we could increase N_4 and simultaneously *change* the bare coupling constants in such a way that the ratio of the correlation length to the linear extension of the universe of volume N_4 (both in terms of lattice units) stayed constant. In the absence of a suitable correlation length, we will use the magnitude of the three-volume fluctuations instead, and identify a *path of constant physics* as a trajectory in phase C along which the discrete four-volume N_4 grows, but the accompanying change in the bare couplings κ_0 and Δ ensures that the three-volume fluctuations likewise increase, in such a way that the ratio between the magnitude of the fluctuations and the mean three-volume stays the same. Fixing this ratio forces us to change bare coupling constants when we increase N_4, in this way tracing out a path that moves toward one of the phase transitions bordering phase C, see Fig. 34.7b for a schematic illustration. Preliminary results from computer simulations to determine the flow defined in this way indicate that it should start quite close to the B–C phase transition if it should resemble the flow line of constant physics shown in the figure, raising again the issue of critical slowing-down near the B–C line.

34.4 Relation to Hořava–Lifshitz Gravity

As described above, our CDT data in phase C can be fitted well to the functional form (34.28), which in turn can be seen as a discretized version of the minisuperspace action (34.27). There is a residual ambiguity in the interpretation of the discrete time coordinate appearing in the identification (34.26), which can be thought of as an overall, finite scaling between the time and spatial directions. As we have emphasized, due to the entropic nature of the effective action, there is no compelling reason to take the geometric length assignments of the regularized theory literally. We have identified the time *coordinate* t with continuum proper time in such a way that we obtain a round four-sphere, which is a perfectly legitimate and physically well-motivated choice. However, as we vary the bare couplings κ_0 and Δ, the overall shape of the computer-generated universe changes in terms of the number of lattice spacings in the time direction relative to those in the spatial directions. Although this change is qualitatively in agreement with the change of α as a function of κ_0 and Δ, there is no detailed quantitative agreement.

Instead of choosing continuum time to be consistent with a continuum S^4-geometry as one moves in phase space, one may be able to find a modified action which describes the observed behavior without performing an overall time rescaling which depends on κ_0 and Δ. This may be especially appropriate in the vicinity of the phase transition, where the length scales one is probing become increasingly Planckian, and one would expect significant contributions to the effective dynamics from terms not contained in the infrared form of the Einstein–Hilbert action including higher order curvature terms.

We will consider yet another generalization, which suggests itself because of the built-in anisotropy between time and space of the CDT set-up, namely, a deformation à la Hořava–Lifshitz [34.68, 69]. A corresponding effective Euclidean continuum action, including measure contributions, and expressed in terms of standard metric variables could be of the form

$$S_H = \frac{1}{16\pi G}\int \mathrm{d}^3x\,\mathrm{d}t N\sqrt{g}\left((K_{ij}K^{ij}-\lambda K^2)+(-\gamma R^{(3)}+2\Lambda+V(g_{ij})\right), \tag{34.35}$$

where K_{ij} denotes the extrinsic curvature and g_{ij} the three-metric of the spatial slices, $R^{(3)}$ the corresponding three-dimensional scalar curvature, N the lapse function, and finally $V(g_{ij})$ a *potential* which in Hořava's continuum formulation would contain higher orders of spatial derivatives, potentially rendering S_H renormalizable. In our case we are not committed to any particular choice of potential $V(g_{ij})$, since we are not imposing renormalizability of the theory in any conventional sense.

An effective $V(g_{ij})$ could be generated by entropy, i. e., by the measure, and may not relate to any discussion of the theory being renormalizable. The kinetic term depending on the extrinsic curvature is the most general such term which is at most second order in time derivatives and consistent with spatial diffeomorphism invariance. The parameter λ appears in the (generalized) DeWitt metric, which defines an ultralocal metric on the classical space of all three-metrics (The value of λ governs the signature of the generalized DeWitt metric

$$G_\lambda^{ijkl} = \tfrac{1}{2}\sqrt{\det g}(g^{ik}g^{jl}+g^{il}g^{jk}-2\lambda g^{ij}g^{kl}),$$

which is positive definite for $\lambda < 1/3$, indefinite for $\lambda = 1/3$ and negative definite for $\lambda > 1/3$. The role of λ in three-dimensional CDT quantum gravity has been analyzed in detail in [34.70, 71].), and the parameter γ can be related to a relative scaling between time and spatial directions. Setting $\lambda = \gamma = 1$ and $V = 0$ in (34.35) we recover the standard (Euclidean) Einstein–Hilbert action.

Making a simple minisuperspace ansatz with compact spherical slices, which assumes homogeneity and isotropy of the spatial three-metric g_{ij}, and fixing the lapse to $N = 1$, the Euclidean action (34.35) becomes a function of the scale factor $a(t)$ (see also [34.72–74], as well as [34.75] for related work in 2+1 dimensions), that is,

$$S_{\mathrm{mini}} = \frac{2\pi^2}{16\pi G}\int \mathrm{d}t\,a(t)^3 \times\left(3(1-3\lambda)\frac{\dot a^2}{a^2}-\gamma\frac{6}{a^2}+2\Lambda+\tilde V(a)\right). \tag{34.36}$$

The first three terms in the parentheses define the IR limit (which in Hořava–Lifshitz gravity is assumed to include a flowing of λ to its *GR value*), while the po-

tential term $\tilde{V}(a)$ contains inverse powers of the scale factor a coming from possible higher order spatial derivative terms.

Our reconstruction of the effective action from the computer data is compatible with the functional form (34.36) of the minisuperspace action. If we were able to extract the constant $\tilde{k}_2$ in front of the potential term in (34.28), it would enable us to fix the ratio $(1-3\lambda)/2\gamma$ appearing in (34.36) [34.76]. At this stage, the precision of our measurements is insufficient to do so. The same is true for our attempts to determine $\tilde{V}(a)$ for small values of the scale factor, which is important for understanding UV quantum corrections to the potential near $a(t) = 0$. Once we have developed a better computer algorithm which allows us to approach the B–C phase transition line more closely, investigating such Planckian properties and testing scenarios of Hořava–Lifshitz type will be within reach.

34.5 Conclusions

In constructing a theory of quantum gravity using causal dynamical triangulations, one of our initial inputs was the Regge action, which appears in the weights of individual spacetimes in the gravitational path integral. However, as we have emphasized repeatedly, the full effective action generated dynamically by performing the nonperturbative sum over histories is only indirectly related to this *bare* action. Likewise, the coupling constant k_1, which appears in front of the effective action and we view as related to the gravitational coupling constant G, has no obvious direct relation to the *bare* coupling κ_0 appearing in the Regge action.

Nevertheless, the leading terms in the effective action for the scale factor are precisely the ones present in (34.27) or, more generally, in the effective Hořava–Lifshitz action (34.36), at least for sufficiently large values of the scale factor. The fact that a kinetic term quadratic in derivatives appears as the leading term in the effective action is perhaps less surprising, but that the correct powers of the (undifferentiated) variable $N_3(i)$ appear in both the kinetic and potential terms in (34.28) is rather remarkable and very encouraging for the entire CDT quantization program.

For the range of bare coupling constants and four-volumes investigated until now our results are compatible with the Einstein–Hilbert action. Better data and more observables will be required to discriminate between a *pure gravity* behavior and an anisotropic deformation à la Hořava–Lifshitz in the deep ultraviolet. A beautiful feature of CDT quantum gravity is that entirely nonperturbative questions of this kind can be formulated explicitly and addressed with the nonperturbative lattice tools available, and – if one is lucky – be answered quantitatively.

References

34.1 S. Weinberg: Ultraviolet divergences in quantum theories of gravitation. In: *General Relativity: Einstein Centenary Survey*, ed. by S.W. Hawking, W. Israel (Cambridge University Press, Cambridge, UK 1979), 790–831

34.2 H. Kawai, M. Ninomiya: Renormalization group and quantum gravity, Nucl. Phys. B **336**, 115 (1990)

34.3 H. Kawai, Y. Kitazawa, M. Ninomiya: Scaling exponents in quantum gravity near two-dimensions, Nucl. Phys. B **393**, 280–300 (1993)

34.4 H. Kawai, Y. Kitazawa, M. Ninomiya: Ultraviolet stable fixed point and scaling relations in $(2+\epsilon)$-dimensional quantum gravity, Nucl. Phys. B **404**, 684–716 (1993)

34.5 H. Kawai, Y. Kitazawa, M. Ninomiya: Renormalizability of quantum gravity near two dimensions, Nucl. Phys. B **467**, 313–331 (1996)

34.6 T. Aida, Y. Kitazawa, H. Kawai, M. Ninomiya: Conformal invariance and renormalization group in quantum gravity near two-dimensions, Nucl. Phys. B **427**, 158–180 (1994)

34.7 M. Reuter: Nonperturbative evolution equation for quantum gravity, Phys. Rev. D **57**, 971–985 (1998)

34.8 A. Codello, R. Percacci, C. Rahmede: Investigating the ultraviolet properties of gravity with a Wilsonian renormalization group equation, Ann. Phys. **324**, 414 (2009)

34.9 M. Reuter, F. Saueressig: Functional renormalization group equations, asymptotic safety, and quantum Einstein gravity (2007), arXiv:0708.1317 [hep-th]

34.10 M. Niedermaier, M. Reuter: The asymptotic safety scenario in quantum gravity, Living Rev. Relativ. **9**, 5 (2006)

34.11 H.W. Hamber, R.M. Williams: Nonlocal effective gravitational field equations and the running of Newton's G, Phys. Rev. D **72**, 044026 (2005)

34.12 D.F. Litim: Fixed points of quantum gravity, Phys. Rev. Lett. **92**, 201301 (2004)
34.13 T. Regge: General relativity without coordinates, Nuovo Cim. **19**, 558 (1961)
34.14 J. Ambjørn, B. Durhuus, J. Fröhlich: Diseases of triangulated random surface models, and possible cures, Nucl. Phys. B **257**, 433–449 (1985)
34.15 J. Ambjørn, B. Durhuus, J. Fröhlich, P. Orland: The appearance of critical dimensions in regulated string theories, Nucl. Phys. B **270**, 457–482 (1986)
34.16 F. David: Planar diagrams, two-dimensional lattice gravity and surface models, Nucl. Phys. B **257**, 45 (1985)
34.17 A. Billoire, F. David: Microcanonical simulations of randomly triangulated planar random surfaces, Phys. Lett. B **168**, 279–283 (1986)
34.18 V.A. Kazakov, A.A. Migdal, I.K. Kostov: Critical properties of randomly triangulated planar random surfaces, Phys. Lett. B **157**, 295–300 (1985)
34.19 D.V. Boulatov, V.A. Kazakov, I.K. Kostov, A.A. Migdal: Analytical and numerical study of the model of dynamically triangulated random surfaces, Nucl. Phys. B **275**, 641–686 (1986)
34.20 B. Dittrich: How to construct diffeomorphism symmetry on the lattice, Proc. 3rd Quantum Gravity Quantum Geom. Sch. (2011)
34.21 V.G. Knizhnik, A.M. Polyakov, A.B. Zamolodchikov: Fractal structure of 2D quantum gravity, Mod. Phys. Lett. A **3**, 819 (1988)
34.22 F. David: Conformal field theories coupled to 2D gravity in the conformal gauge, Mod. Phys. Lett. A **3**, 1651 (1988)
34.23 J. Distler, H. Kawai: Conformal field theory and 2D quantum gravity or Who's afraid of Joseph Liouville?, Nucl. Phys. B **321**, 509 (1989)
34.24 F. David: Loop equations and nonperturbative effects in two-dimensional quantum gravity, Mod. Phys. Lett. A **5**, 1019–1030 (1990)
34.25 J. Ambjørn, J. Jurkiewicz, Y.M. Makeenko: Multiloop correlators for two-dimensional quantum gravity, Phys. Lett. B **251**, 517–524 (1990)
34.26 J. Ambjørn, K.N. Anagnostopoulos: Quantum geometry of 2D gravity coupled to unitary matter, Nucl. Phys. B **497**, 445 (1997)
34.27 J. Ambjørn, K.N. Anagnostopoulos, U. Magnea, G. Thorleifsson: Geometrical interpretation of the KPZ exponents, Phys. Lett. B **388**, 713 (1996)
34.28 J. Ambjørn, J. Jurkiewicz, Y. Watabiki: On the fractal structure of two-dimensional quantum gravity, Nucl. Phys. B **454**, 313–342 (1995)
34.29 H. Kawai, N. Kawamoto, T. Mogami, Y. Watabiki: Transfer matrix formalism for two-dimensional quantum gravity and fractal structures of spacetime, Phys. Lett. B **306**, 19 (1993)
34.30 J. Ambjørn, Y. Watabiki: Scaling in quantum gravity, Nucl. Phys. B **445**, 129 (1995)
34.31 J. Ambjørn, J. Correia, C. Kristjansen, R. Loll: On the relation between Euclidean and Lorentzian 2-D quantum gravity, Phys. Lett. B **475**, 24–32 (2000)
34.32 J. Ambjørn, R. Loll: Non-perturbative Lorentzian quantum gravity, causality and topology change, Nucl. Phys. B **536**, 407–434 (1998)
34.33 C. Teitelboim: Causality versus gauge invariance in quantum gravity and supergravity, Phys. Rev. Lett. **50**, 705–708 (1983)
34.34 C. Teitelboim: The proper time gauge in quantum theory of gravitation, Phys. Rev. D **28**, 297–309 (1983)
34.35 B. Dittrich, R. Loll: Counting a black hole in Lorentzian product triangulations, Class. Quantum Gravity **23**, 3849–3878 (2006)
34.36 P. di Francesco, E. Guitter: Critical and multicritical semi-random $(1+d)$-dimensional lattices and hard objects in d dimensions, J. Phys. A **35**, 897–928 (2002)
34.37 B. Brügmann, E. Marinari: 4-d simplicial quantum gravity with a nontrivial measure, Phys. Rev. Lett. **70**, 1908 (1993)
34.38 S. Bilke, Z. Burda, A. Krzywicki, B. Petersson, J. Tabaczek, G. Thorleifsson: 4-D simplicial quantum gravity: Matter fields and the corresponding effective action, Phys. Lett. B **432**, 279 (1998)
34.39 J. Ambjørn, K.N. Anagnostopoulos, J. Jurkiewicz: Abelian gauge fields coupled to simplicial quantum gravity, J. High Energy Phys. **9908**, 016 (1999)
34.40 J. Laiho, D. Coumbe: Evidence for asymptotic safety from lattice quantum gravity, Phys. Rev. Lett. **107**, 161301 (2011)
34.41 D. Coumbe, J. Laiho: Exploring the phase diagram of lattice quantum gravity, PoS Lattice **2011**, 334 (2011)
34.42 R. Loll: Discrete Lorentzian quantum gravity, Nucl. Phys. Proc. Suppl. **94**, 96 (2001)
34.43 R. Loll: A discrete history of the Lorentzian path integral, Lecture Notes in Physics **631**, 137 (2003)
34.44 J. Ambjørn, J. Jurkiewicz, R. Loll: Quantum gravity, or the art of building spacetime. In: *Approaches to Quantum Gravity*, ed. by D. Oriti (Cambridge Univ. Press, Cambridge 2009) pp. 341–359
34.45 R. Loll: The emergence of spacetime or quantum gravity on your desktop, Class. Quantum Gravity **25**, 114006 (2008)
34.46 J. Ambjørn, A. Görlich, J. Jurkiewicz, R. Loll: The emergence of (Euclidean) de Sitter space-time. In: *Path Integrals – New Trends and Perspectives*, ed. by W. Janke, A. Pelster (World Scientific, Singapore 2008) pp. 191–198
34.47 J. Ambjørn, A. Görlich, J. Jurkiewicz, R. Loll: The quantum universe, Acta Phys. Pol. B **39**, 3309 (2008)
34.48 J. Ambjørn, J. Jurkiewicz, R. Loll: Deriving spacetime from first principles, Ann. Phys. **19**, 186 (2010)
34.49 J. Ambjørn, J. Jurkiewicz, R. Loll: Causal dynamical triangulations and the quest for quantum gravity. In: *Foundations of Space and Time*, ed. by G. Ellis, J. Murugan, A. Weltman (Cambridge Univ. Press, Cambridge 2012)

34.50 J. Ambjørn, J. Jurkiewicz, R. Loll: Lattice quantum gravity – An update, PoS Lattice **2010**, 014 (2010)
34.51 J. Ambjørn, J. Jurkiewicz, R. Loll: Dynamically triangulating Lorentzian quantum gravity, Nucl. Phys. B **610**, 347–382 (2001)
34.52 J. Ambjørn, J. Jurkiewicz, R. Loll: A nonperturbative Lorentzian path integral for gravity, Phys. Rev. Lett. **85**, 924 (2000)
34.53 J. Ambjørn, J. Jurkiewicz, R. Loll: Quantum gravity as sum over spacetimes, Lecture Notes in Physics **807**, 59 (2010)
34.54 J. Ambjørn, A. Görlich, J. Jurkiewicz, R. Loll: CDT – An entropic theory of quantum gravity, arXiv:1007.2560 [hep-th]
34.55 J. Ambjørn, A. Görlich, J. Jurkiewicz, R. Loll: Nonperturbative quantum gravity, Phys. Rep. **519**, 127–210 (2012)
34.56 J. Ambjørn, A. Görlich, S. Jordan, J. Jurkiewicz, R. Loll: CDT meets Hořava–Lifshitz gravity, Phys. Lett. B **690**, 413–419 (2010)
34.57 J. Ambjørn, J. Jurkiewicz, R. Loll: Emergence of a 4D world from causal quantum gravity, Phys. Rev. Lett. **93**, 131301 (2004)
34.58 J. Ambjørn, J. Jurkiewicz, R. Loll: Reconstructing the universe, Phys. Rev. D **72**, 064014 (2005)
34.59 J. Ambjørn, S. Jordan, J. Jurkiewicz, R. Loll: A second-order phase transition in CDT, Phys. Rev. Lett. **107**, 211303 (2011)
34.60 J. Ambjørn, S. Jordan, J. Jurkiewicz, R. Loll: Second- and first-order phase transitions in CDT, Phys. Rev. D **85**, 124044 (2012)
34.61 J. Ambjørn, J. Jurkiewicz, R. Loll: Semiclassical universe from first principles, Phys. Lett. B **607**, 205–213 (2005)
34.62 A. Dasgupta, R. Loll: A proper time cure for the conformal sickness in quantum gravity, Nucl. Phys. B **606**, 357 (2001)
34.63 J. Ambjørn, A. Dasgupta, J. Jurkiewicz, R. Loll: A Lorentzian cure for Euclidean troubles, Nucl. Phys. Proc. Suppl. **106**, 977 (2002)
34.64 J. Ambjørn, J. Jurkiewicz: Four-dimensional simplicial quantum gravity, Phys. Lett. B **278**, 42–50 (1992)
34.65 J. Ambjørn, A. Görlich, J. Jurkiewicz, R. Loll: Planckian birth of the quantum de Sitter universe, Phys. Rev. Lett. **100**, 091304 (2008)
34.66 J. Ambjørn, A. Görlich, J. Jurkiewicz, R. Loll: The nonperturbative quantum de Sitter universe, Phys. Rev. D **78**, 063544 (2008)
34.67 J.B. Hartle, S.W. Hawking: Wave function of the universe, Phys. Rev. D **28**, 2960–2975 (1983)
34.68 P. Hořava: Quantum gravity at a Lifshitz point, Phys. Rev. D **79**, 084008 (2009)
34.69 P. Hořava, C.M. Melby-Thompson: General covariance in quantum gravity at a Lifshitz point, Phys. Rev. D **82**, 064027 (2010)
34.70 T. Budd: The effective kinetic term in CDT, J. Phys. Conf. Ser. **36**, 012038 (2012)
34.71 T. Budd, R. Loll: Exploring torus universe in causal dynamical triangulations, Phys. Rev. D **88**(2), 024015 (2013)
34.72 E. Kiritsis, G. Kofinas: Hořava–Lifshitz cosmology, Nucl. Phys. B **821**, 467 (2009)
34.73 R. Brandenberger: Matter bounce in Hořava–Lifshitz cosmology, Phys. Rev. D **80**, 043516 (2009)
34.74 G. Calcagni: Cosmology of the Lifshitz universe, J. High Energy Phys. **0909**, 112 (2009)
34.75 D. Benedetti, J. Henson: Spectral geometry as a probe of quantum spacetime, Phys. Rev. D **80**, 124036 (2009)
34.76 J. Ambjørn, A. Görlich, J. Jurkiewicz, R. Loll, J. Gizbert-Studnicki, T. Trzesniewski: The semiclassical limit of causal dynamical triangulations, Nucl. Phys. B **849**, 144 (2011)

35. String Theory and Primordial Cosmology

Maurizio Gasperini

String cosmology aims at providing a reliable description of the very early Universe in the regime where standard-model physics is no longer appropriate, and where we can safely apply the basic ingredients of superstring models such as dilatonic and axionic forces, duality symmetries, winding modes, limiting sizes and curvatures, higher dimensional interactions among elementary extended object. The sought target is that of resolving (or at least alleviating) the big problems of standard and inflationary cosmology like the spacetime singularity, the physics of the trans-Planckian regime, the initial condition for inflation, and so on.

35.1 The Standard *Big Bang* Cosmology

In the second half of the last century the theoretical and observational study of our Universe, grounded on one hand on the Einstein theory of general relativity, and on the other hand on astronomical observations of every increasing precision, has led to the formulation (and to the subsequent completion) of the so-called *standard cosmological model* [35.1–3].

Such a model – like all physical models – is based on various assumptions. We should mention, in particular:

i) The assumption that the large-scale spacetime geometry can be foliated by a class of three-dimensional space-like hypersurfaces which are exactly homogeneous and isotropic.
ii) The assumption that the matter and the radiation filling our Universe behave exactly as a perfect fluid with negligible friction and viscosity terms.
iii) The assumption that the radiation is in thermal equilibrium.
iv) The assumption that the dominant source of gravity, on cosmological scales, is the so-called *dark matter* component of the cosmological fluid (invisible, up to now, to all attempted detection procedures of nongravitational type); and so on.

Using such assumptions, the standard cosmological model has obtained a long and impressive series of successes and experimental confirmations, such as:

i) The geometric interpretation of the apparent recessional velocity of distant light sources, together with a precise theoretical formalization of the empirical Hubble law.
ii) The prediction of a relic background of thermal radiation.
iii) The explanation of the process of genesis of the light elements and of the other *building blocks* of our present macroscopic world (like the processes of nucleosynthesis and baryogenesis); and so on.

In spite of these important achievements the standard cosmological model was put in trouble when, in the 1980s, the scientific community started to investigate the problem of the origin of the observed galactic structures, and of the small (but finite) inhomogeneity fluctuations presents in the temperature T of the relic background radiation ($\Delta T/T \approx 10^{-5}$). How did originate the temperature inhomogeneities $\Delta T/T$ and, especially, the matter inhomogeneities $\Delta\rho/\rho$ which are at the grounds of the concentration and subsequent

growth of the cosmic aggregates (cluster of galaxies, stars and planets) that we presently observe? No temperature fluctuation and density fluctuation should exist, on macroscopic scales, if our Universe would be exactly homogeneous and isotropic as required by the standard cosmological model.

This problem was solved by assuming that the standard cosmological model has to be modified, at some very early epoch, by the introduction of a cosmological phase – called *inflation* – characterized by an accelerated expansion rate [35.4–6]. During such a primordial inflationary phase the three-dimensional spatial sections of our Universe underwent a gigantic (almost exponential) growth of proper volume in few units of the Hubbe-time parameter [35.3, 7, 8]. This process was able to amplify the microscopic quantum fluctuations of the matter fields (and of the geometry), thus producing the macroscopic inhomogeneities required for the formation of the matter structures and of the temperature anisotropies we observe today [35.8–10].

A phase of inflationary evolution like that proposed before, however, cannot be extended back in time to infinity (or, to use the standard therminology, cannot be *past eternal* [35.11–13]). If we go back in time to sufficiently earlier epochs we find that the inflationary phase of the standard model has a beginning at a precise instant of time. Before that time, the Universe was in an extremely hot, dense, and curved primordial state – an ultimate concentrate of matter and radiation at extremely high energy and temperature.

This means, in other words, that before starting inflating the Universe was quite close to the so-called *big bang* epoch, namely to the epoch of the huge cosmic explosion which – according to the standard model, even including the inflationary phase – gave rise to the matter and energy species we observe today, and was at the origin of the spacetime itself.

In fact, the big bang epoch of the standard model corresponds (strictly speaking) to a mathematical singularity where the energy density and the spacetime curvature blow up to infinity. We can thus say that to the question *How did the Universe begin?*, the standard cosmological model provides the answer: *The Universe was born from the initial big bang singularity.*

35.1.1 Validity Restrictions of the Standard Cosmological Model

It is well known that standard cosmology is based on the Einstein theory of general relativity, which is a *relativistic* theory of gravity, but *not* a *quantum* theory. Hence, like all classical theories, general relativity has a limited validity range. Because of those limits the standard cosmological model cannot be extrapolated to physical regimes where the energy and the spacetime curvature are too high: this prevents taking too seriously the predictions of such a model about the initial singularity.

We should recall, in fact, that a classical model is valid until the corresponding action $S = Et$ is much larger than the elementary *quantum of action* (or Planck's quantum) h. If we take a cosmological patch of the size given by the Hubble radius c/H, we can then estimate the total involved energy E by multiplying the energy density ρ of the gravitational sources by the spatial volume $(c/H)^3$, containing the contribution of all observable matter and radiation at a given time t. The typical cosmological time scale, on the other hand, is provided by the Hubble time H^{-1}, and the energy density ρ is related to the Hubble time by the Einstein equations, which imply (modulo numerical factors of order one) $\rho = c^2H^2/G$, where G is the Newton gravitational constant. By imposing the condition $Et \gg h$ we then find that the standard cosmological model may give a reliable (classical) description of the Universe provided that

$$\frac{c^5}{GH^2} \gg h. \qquad (35.1)$$

(This condition, in units $h = c = 1$, can also be rephrased as $H \ll M_P$, where $M_P = (hc/G)^{1/2}$ is the Planck mass).

The parameters C, G, and H appearing in the above equation are constant, while the Hubble parameter H is closely related to the spacetime curvature and

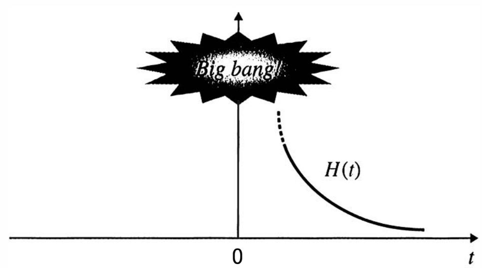


Fig. 35.1 According to the standard cosmological model, the spacetime curvature and the associated Hubble parameter $H(t)$ undergo an unbounded growth as we go back in time, and blow up at the time $t = 0$ of the initial *big bang* singularity

is time dependent, $H = H(t)$. According to the standard model, in particular, H grows as we go back in time, and diverges at the time of the big-bang singularity (Fig. 35.1). Correspondingly, the ratio c^5/GH^2 decreases and goes to zero at the singularity. Hence, before reaching the big bang epoch we necessarily enter the regime where condition (35.1) is violated, and the standard cosmological model is no longer valid.

In order to provide a reliable description of the primordial Universe we should thus use a more general approach, based on a theory able to describe gravity also in the quantum regime. A possible candidate for this theory, which is complete, consistent at all energy scales and, besides gravity, also incorporates all fundamental interactions, is the so-called *theory of strings* [35.14–16].

35.2 String Theory

The name of this theory is due to the fact that it proposes a model where the fundamental *building blocks* of our physical description of nature are one-dimensional extended object (elementary *strings*, indeed), instead of elementary particles. Such strings can be open (of finite length), or closed, and the spectrum of states associated with their vibration modes can reproduce the particle states of the gravitational interaction and of all the other fundamental (electromagnetic, strong and weak) interactions.

In addition, if the string model is appropriately *supersymmetrized* – namely, if we add to each bosonic degree of freedom a corresponding fermionic partner – we arrive at the so-called theory of *superstrings*. This model potentially describes not only all interaction fields, but also their elementary sources (quarks and leptons), and thus all possible species and states of matter [35.14, 15].

But there is more. A basic property of string theory – probably the most revolutionary property, comparing with the other theories – is the property of determining not only the possible form of the interaction terms (which is also done by the usual gauge theories, through the minimal coupling procedure), but also the form of the free-field (kinetic) terms (which in the other theories is always left, to some extent, arbitrary). Indeed, string theory satisfies a new symmetry (called *conformal symmetry*) which rigidly prescribes the allowed free-field dynamics, at any given order of the chosen perturbative expansion [35.14, 15].

Quantizing a string, and imposing that the conformal symmetry is left unbroken by the quantum corrections (i. e. imposing the absence of *conformal anomalies*), one finds, in fact, that – to lowest order – the electromagnetic field *must* satisfy the Maxwell equations, the gravitational field *must* satisfy the Einstein equations, the spinor fields *must* satisfy the Dirac equations, and so on. All field equations, laboriously discovered in the past centuries through the theoretical elaboration of a large amount of empirical data, can be simply *predicted* by string theory even in the absence of any experimental input!

Finally, as already stressed, string theory is valid for all interactions also in the quantum regime, and can thus be used at arbitrarily high energy scales. In particular, unlike general relativity, can be applied to describe the Universe at epochs arbitrarily near to the big bang epoch. In such a limiting high-energy regime the equations we obtain from string theory are different, in general, from the corresponding field-theory equations, and thus it makes sense to ask the question *What's new from string theory about cosmology?*

In particular, *What's new about the very early epochs at the beginning of the Universe?*

35.3 String Cosmology

There are, in particular, two aspects of string theory which can play a relevant role in the formulation of a consistent cosmological scenario.

The first one concerns the so-called *dual symmetry*, typical of one-dimensional extended objects. If such a symmetry is respected (even at the approximate level) by the gravitational dynamics on cosmological scales, then any cosmological phase occurring at $t > 0$, and characterized by a decreasing Hubble parameter H (hence, decreasing curvature), must be associated to a *dual* partner phase, defined at $t < 0$ and characterized by growing H (see [35.17] for a nontechnical illustration of this duality symmetry). It follows in particular that the present cosmological phase, subsequent to the big bang epoch and well described by the standard model, must be preceded in time by another, almost

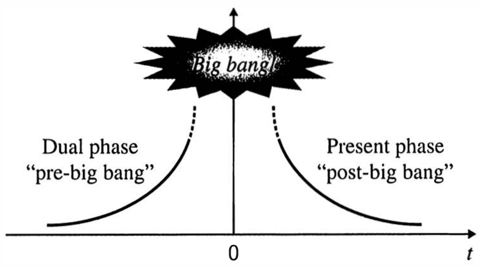


Fig. 35.2 The standard cosmological phase, of *post-big bang* type, is preceded by a (string-theory) dual phase of *pre-big bang* type

specularly symmetric phase, occurring *before the big bang* (Fig. 35.2). Such a duality symmetry should also leave an imprint on the properties of the cosmological perturbations [35.18].

In Fig. 35.2 both phases are characterized by a curvature (and a Hubble parameter) which diverges as t goes to zero. If that would be the case, then the two branches of the cosmological evolution would be causally disconnected by a spacetime singularity, with no chances of merging together into a single coherent model of spacetime evolution. It is here, however, that comes into play another crucial aspect of string theory.

String theory is indeed characterized by a fundamental length λ_s, which is a constant parameter of the string action and which controls the typical size of a quantized string. The physical role played by λ_s is very similar to the role played by the Bohr radius for the atom, which represents the minimal allowed size of the quantum electronic orbitals. The numerical value, however, is quite different: we may expect in fact $\lambda_s \approx 10^{-33}$ cm (i. e. a value of λ_s which is about 10 times that of the Planck length $\lambda_P = h/M_P c^2$), in order that string theory may include a realistic description of all fundamental interactions (different values of λ_s are possible in the presence of large extra dimensions, see below).

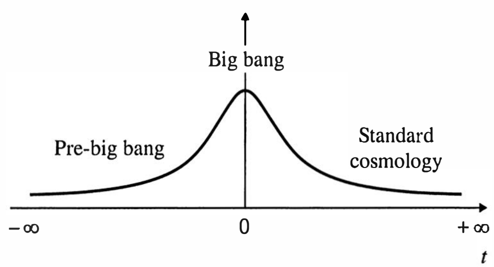


Fig. 35.3 Time evolution of the curvature scale and of the energy density in a typical example of string-cosmology scenario. The big bang epoch does not correspond to a singularity (like in the standard model) but to a phase of maximal, finite curvature. The Universe evolves starting from a flat, cold and empty state called the *string perturbative vacuum*, asymptotically localized at $t = -\infty$

Aside from the particular numerical value of λ_s, what is important, also, is that proper distances and sizes smaller than λ_s have no physical meaning in a string-inspired model. It follows that, in a string-cosmology context, the Hubble radius c/H has to be constrained by the condition $c/H \gtrsim \lambda_s$. Since the Hubble radius is directly related to the inverse of the spacetime curvature, we can deduce that the curvature cannot blow up to infinity, because of the constraint $H \lesssim c/\lambda_s$. Hence, when a given spacetime region has reached the limiting value $H \approx c/\lambda_s$, its geometrical state can only evolve in two ways: it can either stabilize at such a maximum value, or start decaying toward lower curvature states after a bounce induced by appropriate *stringy* effects [35.19].

In such a context, the big-bang singularity predicted by the standard model and sharply localized at a given epoch (say, $t = 0$), has thus to be replaced by an extended phase of very high (but finite) maximal curvature: the so-called *string phase* [35.20, 21]. By combining the existence of the dual symmetry and of a minimal length scale, a string-based model can thus complete the standard cosmological scenario by removing the curvature singularity and extending the physical description of the Universe back in time, beyond the big bang, to infinity. The *big bang era* is still there, but it is deprived of the standard role of initial singularity: it corresponds, instead, to the epoch marking the transition between the growing curvature and the decreasing curvature regime (Fig. 35.3).

Within such a cosmological scenario (first presented in detail in [35.22]) the initial cosmological state is no longer localized at $t = 0$, but it is moved to the limit $t \to -\infty$, and corresponds to an asymptotic state usually called the *string perturbative vacuum*. Such a new initial state, as illustrated in Fig. 35.3, turns out to be a sort of specularly symmetric version of the final state that would be reached in the asymptotic future by a Universe which keeps expanding for ever according to the standard cosmological dynamics. Namely, a flat, empty and cold initial state, drastically different from the initial hot, explosive state, extremely curved and concentrated, proposed by the standard scenario.

There is, however, a possible asymmetry between the initial and final state of the above string-cosmology

model, due to the coupling strength of the fundamental interactions: such a coupling tends to zero as $t \to -\infty$, while it may become very strong in the opposite limit, if not appropriately stabilized [35.20, 21]. This growth of the coupling can be accompanied, in principle, also by a large amount of entropy production [35.23].

35.4 A Higher Dimensional Universe

String theory, which is at the grounds of the cosmological scenario described in the previous section, can be consistently formulated only in the context of a higher-dimensional spacetime manifold.

In fact, in order to consistently quantize a bosonic string without introducing *ghosts* (states of negative norm), and without violating the Lorentz symmetry, one must introduce a generalized spacetime manifold with 26 dimensions [35.14, 15]. In this way, however, one obtains a model which has still a pathology, as it contains *tachyons* (states of imaginary mass), which we believe should be absent in any realistic physical model.

In order to eliminate the tachyons, we can generalize the bosonic string model by adding fermion states and considering the so-called *superymmetric* string models, or *superstring* models. In that case, a consistent quantization requires 10 spacetime dimensions, and the number of total dimensions is to be increased up to 11 (with one time-like and 10 space-like dimensions) if we require that the five possible types of superstrings may be connected by duality transformations, and may represent various weak-coupling regimes of a more fundamental theory, called *M-theory* [35.24, 25].

Hence, whatever string model is assumed to apply, it is clear that the associated string cosmology scenario must be referred to a higher dimensional Universe. On the other hand, all present phenomenological experience (including the most sensitive high-energy experiments) points at a world with one time-like and *only three* space-like dimensions. We are thus naturally led to the following questions:

> *If string theory is correct, and the Universe in which we live has a number $d > 3$ of spatial dimensions, why our experience is only limited to a three-dimensional space? why we cannot detect the additional* extra *dimensions? what happened to those dimensions, if they really exist?*

There are at least two possible answers to the above questions.

There is an *old-fashioned* answer – which, for a long time, has been also the only possible answer to the previous questions – dating back to the so-called Kaluza–Klein model, formulated at the beginning of the last century [35.26, 27] in the context of a higher dimensional version of general relativity. According to such model, we cannot detect the extra dimensions because such dimensions are compactified on length scales of the extremely small size (hence, they need extremely high energies in order to be experimentally resolved).

We can take, as a simple example, a long and very thin cylinder. A cylinder is a two-dimensional object but, if it is observed from a distance much larger than its radius, it may appear (in all respects) as being one-dimensional, extended in length but deprived of any sensible thickness. In the same way the spatial extension of our Universe could be largely asymmetric, with three spatial dimensions macroscopically expanded on a large scale, while all the other dimensions *rolled up* in a highly compact way, and confined on a very small length scale – of order (for instance) of λ_s. If we do not have a sufficiently powerful instrument, able to resolve the required (very tiny) distance scales, we will always observe only three spatial dimensions.

Very recently, however, a new possible answer to the dimensionality problem has been suggested by theoretical studies mainly performed in the second half of the 1990s, and closely related to particular string-model configurations, called *branes* [35.25]. Such a new answer states that we cannot *see* the extra dimensions simply because the fundamental interactions propagate only along three spatial dimensions. All instruments we use to explore the world around us (starting from our eyes up to the more powerful and sophisticated technological tools) have indeed a working mechanism based on the fundamental (electromagnetic, nuclear, and so on) interactions. If such interactions are living only on a restricted subspace of the full spacetime manifold (like, for instance, waves which propagate on the surface of a pond, and not in the direction orthogonal to the pond surface), then the extra dimensions are hidden to our direct experience, even if they are largely (or infinitely) extended.

This second possible answer to the dimensionality problem has suggested new, interesting types of cosmological models, formulated in the context of the so-called *brane-world* scenario [35.28].

35.5 Brane Cosmology

According to the so-called brane-world cosmology, our Universe could be a four-dimensional *slice* of a higher dimensional *bulk* manifold. The elementary charges sourcing the gauge interactions are confined on a three-dimensional hypersurface Σ_3 associated to an object called *Dirichlet* 3-*brane* (or D_3-brane), and we cannot detect the external spatial dimensions because the gauge fields of those charges can propagate only on the *world-volume* Σ_4 swept by the time evolution of the brane. (It should be recalled that the description of our Universe as a four-dimensional *domain wall* embedded in a higher dimensional bulk spacetime was previously suggested, with different motivations, also in [35.29]).

In a string theory context, however, the confinement mechanism is not equally efficient for all fundamental interactions. Gravity, in particular, is not at all (or only partially [35.30]) confined, so that it can propagate even outside the brane spacetime. This possibility is illustrated in Fig. 35.4, which shows a brane spacetime Σ_4 with two possible sources of interactions. One is a charge, source of the electromagnetic field: the associated electromagnetic waves (or photons) are strictly confined to propagate only on Σ_4. The other is a mass, source of the gravitational field: the associated gravitational waves (or gravitons) can leave the brane spacetime Σ_4 and propagate through the external spatial dimensions.

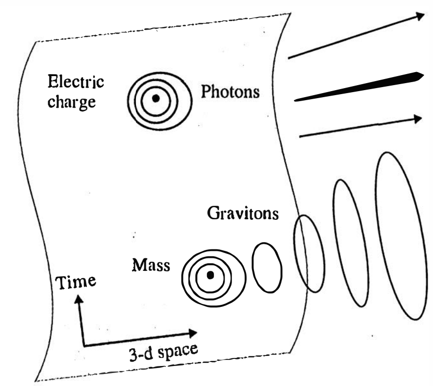


Fig. 35.4 A brane-Universe with one time-like and three space-like dimensions, embedded in an external bulk spacetime characterized by six (according to superstring theory) or seven (according to M-theory) extra spatial dimensions. Electromagnetic forces are confined on the brane spacetime, while gravitational forces propagate also in the directions external to the brane

This property of the gravitational field is quite important because, if the higher dimensional bulk spacetime contains two (or more) fundamental branes, they can interact among themselves gravitationally. And this possibility leads us to an interesting geometric interpretation of the big bang mechanism, namely of the high-energy process which has marked the beginning of the standard cosmological phase, bringing the Universe to the form we are presently observing.

In fact, during the high-curvature phase localized around $t = 0$ (Fig. 35.3), the Universe, if higher dimensional, tends to be filled by branes which are spontaneously produced in pairs from the high-energy vacuum, and which can gravitationally (and strongly) interact among themselves [35.21]. According to string theory, on the other hand, the total gravitational force in a higher dimensional spacetime includes various components: we should mention, in particular, the symmetric-tensor contribution associated to the *graviton*, the scalar contribution associated to the *dilaton*, and the antisymmetric-tensor contribution associated to the *axion* [35.20, 21].

The first two types of forces are always attractive, while the axion force is repulsive between sources of the same sign and attractive between sources of opposite sign (like, for instance, a brane and an anti-brane, characterized by opposite axionic charges). It follows, in particular, that if we have two identical branes (or anti-branes) in an initial static and symmetric state, then the axion repulsion exactly cancels the attraction due to the graviton and to the dilaton, and the net result-

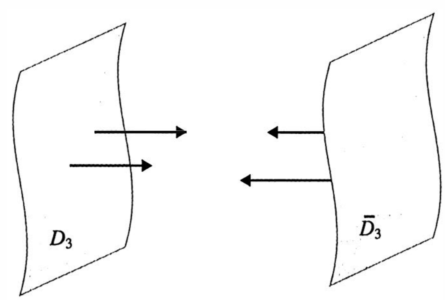


Fig. 35.5 A brane (D_3) and an anti-brane ($\overline{D}_3$) tend to collide because the mutual gravitational force they experience in a higher dimensional spacetime is always of attractive type (like the electric force acting between a charged particle and the corresponding antiparticle in the usual three-dimensional space)

ing force is vanishing. If we have instead a brane and an anti-brane then the total gravitational force is always nonvanishing and attractive, quite irrespective of their initial configuration.

Because of such attractive force branes and anti-branes, copiously produced during the high-energy pre-big bang phase, tend to collide among themselves (Fig. 35.5): it could be, therefore, that it was the collision of our brane-Universe with an anti-brane to simulate the big bang explosion, and trigger the transition from the pre-big bang phase to the phase of standard (post-big bang) evolution. This type of scenario is very similar to the so-called *ekpyrotic* model (first proposed in [35.31] and later embedded in the context of a more general type of *cyclic* cosmologies [35.32]), with the only difference that, in the ekpyrotic case, the 3-branes are *domain walls* representing the spacetime boundaries.

35.6 Conclusion

String theory, M-theory, and the related models of brane interactions suggests new and interesting scenarios for the birth of the Universe and its subsequent primordial evolution, not necessarily limited in time by a big-bang singularity. They can be tested by present (or near-future) observations concerning the properties of the cosmic background of relic gravitational radiation [35.33] and of the so-called dark energy (or quintessence field) dominating the large-scale dynamics [35.34, 35].

Some of those scenario have been briefly introduced and illustrated in the previous sections. But there are also other, equally interesting scenarios closely related to the previous ones, among which I would like to mention the string-gas [35.36] and brane-gas [35.37] cosmologies, based on the repulsive mechanism of winding modes, as well as more general bouncing cosmology models [35.38–40]. Also, models of brane anti-brane inflation [35.41, 42], where the (time-varying) distance between the two branes plays the role of the inflation field.

All these models have many (and interesting) phenomenological implications, but – as usual in a cosmological context – many studies and many observational data are required before being able of selecting the model most appropriate to our Universe. Thus, we can easily predict that we still have in front of us many years of work and – maybe – of surprising findings.

References

35.1 S. Weinberg: *Gravitation and Cosmology* (Wiley, New York 1972)

35.2 S. Dodelson: *Modern Cosmology* (Academic Press, San Diego 2003)

35.3 S. Weinberg: *Cosmology* (Oxford Univ. Press, Oxford 2008)

35.4 A. Guth: The inflationary universe: A possible solution to the horizon and flatness problems, Phys. Rev. D **23**, 347 (1981)

35.5 A.D. Linde: A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems, Phys. Lett. B **108**, 389 (1982)

35.6 A. Albrecht, P.J. Steinhardt: Cosmology for grand unified theories with radiatively induced symmetry breaking, Phys. Rev. Lett. **48**, 1220 (1982)

35.7 E.W. Kolb, M.S. Turner: *The Early Universe* (Addison Wesley, Redwod City 1990)

35.8 A.R. Liddle, D.H. Lyth: *Cosmological Inflation and Large-Scale Structure* (Cambridge Univ. Press, Cambridge 2000)

35.9 V.F. Mukhanov: *Physical Foundation of Cosmology* (Cambridge Univ. Press, Cambridge 2005)

35.10 R. Durrer: *The Cosmic Microwave Background* (Cambridge Univ. Press, Cambridge 2008)

35.11 A. Borde, A. Vilenkin: Eternal inflation and the initial singularity, Phys. Rev. Lett. **72**, 3305 (1994)

35.12 A. Borde, A. Guth, A. Vilenkin: Inflationary spacetimes are incomplete in past directions, Phys. Rev. Lett. **90**, 151301 (2003)

35.13 A. Mithani, A. Vilenkin: Did the universe have a beginning?, arXiv:1204.4658

35.14 M.B. Green, J. Schwartz, E. Witten: *Superstring Theory* (Cambridge Univ. Press, Cambridge 1987)

35.15 J. Polchinski: *String Theory* (Cambridge Univ. Press, Cambridge 1998)

35.16 B. Zwiebach: *A First Course in String Theory* (Cambridge Univ. Press, Cambridge 2009)

35.17 M. Gasperini: *The Universe Before the Big Bang: Cosmology and String Theory* (Springer, Berlin, Heidelberg 2008)

35.18 R. Brustein, M. Gasperini, G. Veneziano: Duality in cosmological perturbation theory, Phys. Lett. B **431**, 277 (1998)
35.19 M. Gasperini, M. Giovannini, G. Veneziano: Cosmological perturbations across a curvature bounce, Nucl. Phys. B **694**, 206 (2004)
35.20 M. Gasperini: *Elements of String Cosmology* (Cambridge Univ. Press, Cambridge 2007)
35.21 M. Gasperini, G. Veneziano: The Pre-big bang scenario in string cosmology, Phys. Rep. **373**, 1 (2003)
35.22 M. Gasperini, G. Veneziano: Pre-big bang in string cosmology, Astropart. Phys. **1**, 317 (1993)
35.23 M. Gasperini, M. Giovannini: Quantum squeezing and cosmological entropy production, Class. Quantum Gravity **10**, L133 (1993)
35.24 E. Witten: String theory dynamics in various dimensions, Nucl. Phys. B **443**, 85 (1995)
35.25 K. Becker, M. Becker, J.E. Schwarz: *String Theory and M-Theory: A Modern Introduction* (Cambridge Univ. Press, Cambridge 2006)
35.26 T. Kaluza: On the problem of unity in physics, Sitzungsber. Preuss. Akad. Wiss. Berlin **1921**, 966 (1921)
35.27 O. Klein: Quantum theory and five-dimensional theory of relativity, Z. Phys. **37**, 895 (1926)
35.28 D. Langlois: Brane cosmology: An introduction, Prog. Theor. Phys. Suppl. **148**, 181 (2003)
35.29 V.A. Rubakov, M. Shaposhnikov: Do we live inside a domain wall?, Phys. Lett. B **125**, 136 (1983)
35.30 L. Randall, R. Sundrum: An alternative to compactification, Phys. Rev. Lett. **83**, 4960 (1999)
35.31 J. Khoury, B.A. Ovrut, P.J. Steinhard, N. Turok: The Ekpyrotic universe: Colliding branes and the origin of the hot big bang, Phys. Rev. D **64**, 123522 (2001)
35.32 P.J. Steinhardt, N. Turok: Cosmic evolution in a cyclic universe, Phys. Rev. D **65**, 126003 (2002)
35.33 M. Gasperini: Elementary introduction to pre-big bang cosmology and to the relic graviton background. In: *Gravitational Waves*, ed. by I. Ciufolini, V. Gorini, U. Mischella, P. Frè (IOP Publishing, Bristol 2001) pp. 280–337
35.34 M. Gasperini: Dilatonic interpretation of the quintessence?, Phys. Rev. D **64**, 043510 (2001)
35.35 L. Amendola, M. Gasperini, F. Piazza: Fitting type Ia supernovae with coupled dark energy, J. Cosmol. Astropart. Phys. **0409**, 014 (2004)
35.36 R. Brandenberger, C. Vafa: Superstrings in the early universe, Nucl. Phys. B **316**, 391 (1989)
35.37 S. Alexander, R. Brandenberger, D. Easson: Brane gases in the early universe, Phys. Rev. D **62**, 103509 (2000)
35.38 T. Biswas, A. Mazumdar, W. Siegel: Bouncing universes in string-inspired gravity, J. Cosmol. Astropart. Phys. **0603**, 009 (2006)
35.39 T. Biswas, T. Koivisto, A. Mazumdar: Towards a resolution of the cosmological singularity in non-local higher derivative theories of gravity, J. Cosmol. Astropart. Phys. **1011**, 008 (2010)
35.40 Y.-F. Cai, D.A. Easson, R. Brandenberger: Towards a nonsingular bouncing cosmology, J. Cosmol. Astropart. Phys. **1208**, 020 (2012)
35.41 C. Burgess, M. Majumdar, D. Nolte, F. Quevedo, G. Rajesh, R.J. Zhang: The Inflationary brane antibrane universe, J. High Energy Phys. **0107**, 047 (2001)
35.42 S. Kachru, R. Kallosh, A. Linde, J. Maldacena, L. McAlliste, S.P. Trivedi: Towards inflation in string theory, J. Cosmol. Astropart. Phys. **0310**, 013 (2003)

36. Quantum Spacetime

Carlo Rovelli

The recent progress towards the construction of a quantum theory of gravity has been impressive, in particular thanks to the Fairbairn–Meusburger–Han theorems on the finiteness of the spinfoam expansion [36.1, 2], and the Freidel-Conrady–Barrett et al.–Han theorems on its classical limit [36.1, 3–5]. This advance yields a very good understanding of how quantum spacetime can be described. I summarize the result of these developments, focusing on the conceptual aspects of the problem: the emerging nature of quantum spacetime, and the revision of the concepts of space and time it demands.

36.1 General Ideas for Understanding Quantum Gravity

The theoretical side of the problem of quantum gravity is relatively well defined: write a quantum theory without uncontrollable divergences whose classical limit is general relativity, with its matter couplings. This problem should not be confused with that of unification, which is independent, and probably unrelated. Quantum chromodynamics (QCD) is a good quantum theory of the strong interactions, not a unification with other interactions.

In a quantum theory of gravity, spacetime can no longer be thought as a four-dimensional manifold, for the same reason for which a quantum particle does not have a trajectory. This forces us into a full revision of the notions of time and space. Therefore to understand quantum gravity, we have to start by getting rid of the idea that space is a 3-D metric space, time is a one-dimensional flowing something, or spacetime is a differentiable manifold. We must replace these conceptual tools with others, compatible with a quantum theory of spacetime.

Here I describe the new conceptual tools that may work to understand quantum spacetime.

36.2 Time

In nonrelativistic physics, we describe change in terms of evolution with respect to an external time variable, ideally measured by a clock dynamically independent from the system under consideration. This clock defines the independent variable t in terms of which the dynamics of the dependent variables $q_n(t)$ that describe the system is given.

In general relativistic physics, this formal structure does not work anymore. Instead, we must include the independent parameter of the evolution among the other variables of the system, where it is on the same footing as the other variables and generically indistinguishable from the others. (Because no clock can be decoupled from the gravitational field.) Accordingly, physics does not anymore describe the evolution of the variables *in time*, but rather the *relative* evolution of the variables, namely the evolution of the variables with respect to one another.

The conceptual step is analogous to the step taken by describing a curve in the (x, y) plane in terms of a relation $f(x, y) = 0$ rather than in the form $y = y(x)$. The first option is clearly more general than the second.

In the canonical language, this means that we must work with an $(N+1)$-dimensional *extended* configuration space $\mathcal{E}$, if N is the number of degrees of freedom, and the dynamics is not determined by a Hamiltonian, but by a Hamiltonian constraint C on the corresponding phase space.

The dynamics of a finite-dimensional system is compactly captured by the *Hamilton function*, which is a function on $\mathcal{E} \times \mathcal{E}$ defined as the value of the action on a solution of the equation of motion interpolating between two given points in $\mathcal{E}$. For instance, the free particle dynamics is captured by the Hamilton function

$$S(x, t; x', t) = \frac{(x-x')^2}{2m(t-t')} . \tag{36.1}$$

The derivatives of S with respect to the two variables x and t (treated on equal footing) give the two momenta p_x and p_t and these satisfy the Hamiltonian constraint

$$C(x, t, p_x, p_t) = p_t + \frac{p_x^2}{2m} = 0 . \tag{36.2}$$

In other words, the Hamilton function satisfies the Hamilton–Jacobi equation

$$C\left(x, t, \frac{\partial S}{\partial x}, \frac{\partial S}{\partial t}\right) = 0 . \tag{36.3}$$

The quantum version of this equation,

$$C\left(x, t, \frac{\partial}{\partial x}, \frac{\partial}{\partial t}\right) \Psi(x, t) = 0 , \tag{36.4}$$

is called the Wheeler–DeWitt equation. It reduces to the time-dependent Schrödinger equation in nonrelativistic systems, when we single out one coordinate on $\mathcal{E}$ as the time variable. But its validity is more general.

A quantum theory is also defined by its transition amplitudes, which determine the relative probability of different *processes*. If spectra are continuous, transition amplitudes are functions on $\mathcal{E} \times \mathcal{E}$, like the Hamilton function. In fact, in the approximation where the Planck constant $\hbar$ can be considered small, the transition amplitude satisfies

$$W \approx e^{\frac{i}{\hbar} S} . \tag{36.5}$$

When everything is well defined, the transition amplitudes are the matrix elements of an operator $P \approx \delta(C)$ that *projects* on the solutions of the Wheeler–DeWitt equation. (If C has a continuous spectrum, this equation is properly defined using a Gelfand triple, or equivalent strategies. See [36.6] for a full discussion.)

$$W(x, t; x', t) = \langle x, t | P | x', t' \rangle . \tag{36.6}$$

The transition amplitudes can also be defined à la Feynman, by a functional integral over paths going from the first to the second point in $\mathcal{E}$, weighted by the exponential of $\frac{i}{\hbar}$ times the classical action of the path. In this representation, it is immediate to see why (36.5) holds: it is just the saddle point approximation of the path integral.

The dynamics of a quantum theory is defined by giving the Wheeler–DeWitt equation (the Schrödinger equation in the nonrelativistic case.), *or* by giving the path-integral representation of the transition amplitudes, *or* by directly giving the transition amplitudes, perhaps in terms of a perturbative expansion. The last option is that of the Feynman-rules definition of quantum electrodynamics (QED) and is also the option used below for quantum gravity. This option circumvents the mathematical complications of the Wheeler–DeWitt equation or the full path integral.

The quantum dynamics determines the probability of the state $|x, t\rangle$, given the state $|x', t'\rangle$. Notice that the quantum states of the theory live on the *boundary* of the *process* considered.

This formal structure can be generalized to a field theory as follows. Consider a closed compact region in spacetime and let Σ be its boundary (Fig. 36.1). The states of the theory live on Σ and describe the possible outcomes of any interaction at the boundary. (Or *measurement*, in standard parlance. But *measurement* has a badly misleading connotation: it sounds as to refer to the presence of a human being, or a recording device.

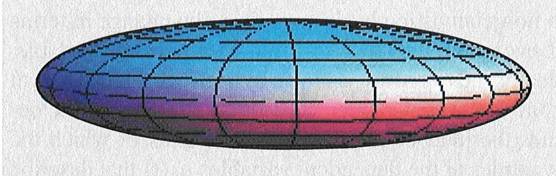


Fig. 36.1 The state described by a spin network can be taken to give the geometry of the three-dimensional hypersurface surrounding a finite 4-D spacetime region

Nothing of this is required to make sense of quantum theory [36.7].)

The quantum amplitudes are functions of boundary data and determine the probability of a process defined by a given boundary state. The boundary states define a Hilbert space $\mathcal{H}$ and the transition amplitudes can be thought, à la Feynman, as the path integral of the field in the bulk. (The boundary does not need to be considered split a priori into past and future. If we do consider this split, then the boundary Hilbert space splits into the tensor product of an *in* and an *out* Hilbert spaces. Tensor product states correspond to pure states, while generic states include statistical states [36.8].)

In quantum gravity, the transition amplitudes describe the full process in the bulk including the gravitational phenomena; therefore, there is no Riemannian manifold inside, as there is no trajectory of a quantum particle. The boundary data include the gravitational boundary data, and these amount to a specification of boundary *metric* quantities. Therefore information such as *the time lapsed* during the process or *the physical distance* between two boundary points is not specified externally: they are already contained in the boundary data about the boundary value of the gravitational field, that is, the metric. This is the beautiful and subtle manner in which time and space are reinterpreted in quantum gravity: as gravitational properties of the boundary data for *physical processes*.

Note that in (36.6) the arguments of the transition amplitudes are coordinates on $\mathcal{E}$ only if the quantum spectrum of these is continuous. If it is discrete, the arguments of the transition amplitudes are the quantum numbers labeling the discrete spectrum. This is what will happen below.

Summarizing, to construct a quantum theory of gravity we need two ingredients: a boundary Hilbert space capable of describing the possible outcome of interactions with (*measurements* of) the gravitational field and matter on the boundary of a process, and transition amplitudes for any given boundary states. The theory will have the appropriate classical limit if the transition amplitudes behave as (36.5) for small $\hbar$, where S is the Hamilton function of general relativity.

Both these ingredients are constructed below.

This structure circumvents entirely the so-called *problem of time* of quantum gravity. The problem of time is resolved by this way of defining the quantum dynamics. The theory is about probabilities assigned to alternative processes. We may avoid talking about *time* altogether, we may forget the word *time*, and still fully and consistently describe change in the world.

36.3 Infinities

Classical field theories have an infinite number of degrees of freedom. Quantum field theories that describe the world well, such as QED and QCD, are commonly defined by building a quantization of a system with a finite number of degrees of freedom, namely quantizing a *truncated* system, and then studying the limit where the truncation is refined. In QED one defines the n-particle Hilbert space, and then formally define the Fock space as an appropriate limit when n increases. When using Feynman graphs, we compute a finite number of graphs, say up to n vertices, and, most of the times, just pay a lip service to the limit where n is increased. In lattice QCD, we define the theory on a finite lattice, and then study the behavior of physical quantities as the number of lattice cells increases.

Quantum gravity also needs to be defined using a truncation.

However, the way the continuous limit is recovered is peculiar, because of the pattern of dimensions in the theory. Conventional interacting quantum field theories suffer for ultraviolet divergences. These come from the dynamics of arbitrarily short-scale (high-momentum) modes. To deal with these, an artificial cut-off is introduced. This is then sent to zero, to get rid of its artificial effects. In perturbative QED and QCD, we consider the limit where the number of vertices n is increased *and* the cut off is removed. In lattice QCD, we study the behavior of physical quantities as the number of lattice cells increases *and* the lattice spacing is taken to zero.

In quantum gravity, on the other hand, there is a scale in the theory, which is the Planck scale $\sqrt{\hbar G}$. A large indirect evidence indicates that this scale determines the maximal physical scale for the field modes. Intuitively, a higher mode falls into its own black hole, because its high energy density generate horizons smaller than its wavelength. Therefore in quantum gravity there is an *intrinsic* cut off, *already built in* into the theory. This implies that there is no artificial cut off that needs to be taken to infinity, in constructing the theory. Unlike QCD, quantum gravity can be defined by first truncating the classical theory and then increasing the lattice size *without* taking a cut off to zero.

See [36.9] for a detailed discussion of the source of the difference.

This can also be seen from a different perspective. In lattice QCD, the size of a cell is an external parameter that enters in the definition of the dynamics. We must tune it to zero to get to the physical theory. In gravity, the physical size of a cell is precisely the dynamical field itself, because the gravitational field is the metric. In any transition amplitude, it is determined by the boundary data. Therefore there is no parameter to tune. Concretely, the Wilson action of given lattice (which defines a truncation of QCD) depends on a parameter, the Regge action on a given triangulation (which defines a truncation of general relativity) does not.

In conventional high-energy physics, the need for the removal of the cut off implies that the theory must behave as if it was in the large-scale limit with respect to the cut-off scale. To avoid triviality, this implies that the theory must sit on a critical point. The same is not true in quantum gravity for the reason explained. Conventional quantum field theory is like condensed matter on critical points, where the theory becomes independent from the atomic scale; quantum gravity is like condensed matter *away* from critical points, where the atomic scale is not lost and determines the macroscopic parameters.

The *atomic* scale in quantum gravity is the Planck scale. It affects large-scale physics in the same manner in which the Bohr radius affects the normal physics of a piece of matter.

This is the key structural difference between quantum gravity and conventional quantum field theory.

36.4 Space

The Hilbert space $\mathcal{H}_\Gamma$ that represents the gravitational field on a given boundary, at fixed truncation of the theory, can be defined as follows. Let Γ be an oriented graph (defined solely by its combinatorial structure). Intuitively, Γ is the dual graph of a cellular decomposition of Σ (Fig. 36.2). The graph determines the truncation. Refining the graph leads to a better approximation of the theory.

Associate with each link l of the graph an $SU(2)$ element U_l. The states of the theory are given by square integrable functions $\psi(U_l)$ invariant under the gauge transformations

$$\psi(U_l) \mapsto \psi(\Lambda_{s(l)} U_l \Lambda_{t(l)}^{-1}) , \quad (36.7)$$

where $s(l)$ and $t(l)$ are the two nodes where the link l starts and ends and $\Lambda_n \in SU(2)$ for any node n of the graph. Square integrability is under the $SU(2)$ Haar measure.

The remarkable feature of this Hilbert space is that it describes a 3-D curved metric geometry in the classical limit. In a sense, this should not be surprising: the main result of Ashtekar's formulation of general relativity [36.10] is that gravity can be described using the phase space of an $SU(2)$ Yang–Mills theory, and the Hilbert space $\mathcal{H}_\Gamma$ is precisely the Hilbert space of a lattice $SU(2)$ Yang–Mills theory.

Therefore, it describes precisely the gravitational field on the boundary of a process [36.11–13]. More precisely, at each given truncation Γ, semiclassical states describe the discrete geometry of a piecewise flat cellular decomposition of a curved *metric* space.

This important result is described in detail in many reviews of loop quantum gravity. See for instance [36.14]. Here I give only a sketchy summary. The key point is that there are natural derivative operators defined on the Hilbert space $\mathcal{H}_\Gamma$: the left-invariant vector fields at the nodes, along the links. These can be shown to satisfy algebraic properties that imply that they are in one-to-one correspondence with the quantities describing the metric geometry of a discrete space. In particular, the Casimir on each link is the area of the corresponding face bounding two cells. The scalar product of two links emerging from the same node determines the angle between the normals of the cor-

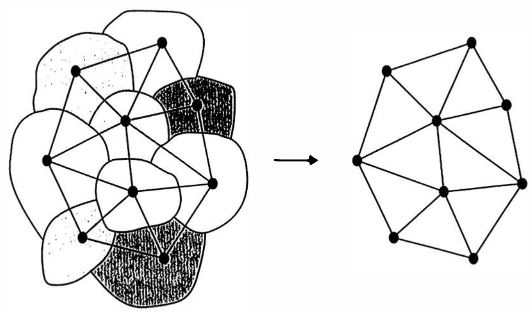


Fig. 36.2 The graph Γ is the dual of a cellular decomposition of the 3-D boundary of the process

responding faces. Thus the expectation values of these operators define a piecewise flat geometry on the cellular decomposition (Fig. 36.3).

A complete set of commuting observables in $\mathcal{H}_\Gamma$ is provided by the areas of the faces and the volume of the 3-D cells. Accordingly, the Hilbert space admits a basis $|\Gamma, j_l, v_n\rangle$, called the spin network basis [36.15], labeled by three groups of quantum numbers: the graph Γ itself, which gives the connectivity of the cells, the spins j_l associated with the faces, that are the quantum numbers of the areas and, at each node n, the quantum number v_n of the volume of the corresponding cell. Since areas of surfaces and volume of cells do not fully determine the classical geometry, the rest of the geometry fluctuates. This is a situation analogous to angular momentum theory, where only L^2 and L_z can be diagonalized simultaneously.

An important result is that the spectrum of area and volume is discrete [36.16, 17]. This is the realization of the intuitive idea of the existence of a physical cut off at the Planck scale. Intuitively, the physical size of the polyhedra of Fig. 36.3 can never become *smaller than the Planck size*. This is a typical quantum phenomenon: the value of the angular momentum can never become smaller than $\hbar/2$. In quantum gravity, it is the reason of the ultraviolet finiteness of the transition amplitudes.

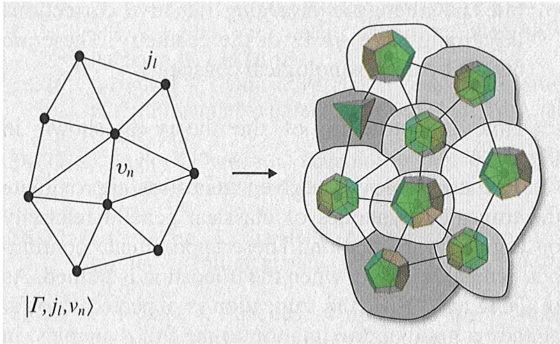


Fig. 36.3 A spin network and the *quanta of space* it describes

36.4.1 Transition Amplitudes

The transition amplitude associated with a given boundary state is defined as follows. First, pick a two-complex C bounded by Γ. Call f its faces, e its edges, and v its vertices. (For simplicity, I consider here only two complexes that are dual to a 4-D triangulation [36.18].). Then

$$W_C(U_l) = \int_{SU(2)} \mathrm{d}h_{vf} \prod_f \delta\left(\prod_v h_{vf}\right) \times \prod_v A_v(h_{vf}) , \tag{36.8}$$

where the vertex amplitude is

$$A(h_f) = \int_{SL(2,C)} \mathrm{d}g_e \prod_f \sum_j d_j Tr_j \times \left[h_f Y^\dagger g_e g_{e'} Y\right] . \tag{36.9}$$

$d_j = 2j + 1$ and Y maps the $SU(2)$ representation of spin j into the spin j subspace of the $SL(2,C)$ unitary representation determined by the discrete spin $k = j$ and the continuous parameter $p = \gamma j$. The parameter γ is a free parameter in the theory. (For a recent clarifying discussion on the Hamiltonian structure of the theory, see [36.19].). This is the full definition of a quantum theory of gravity.

This amplitude was derived in [36.20] building on [36.21–24], and is sometimes called the EPRL (Engle-Pereira-CR-Livine) amplitude. For details, full references, and a derivation of these expressions from the classical action of GR see [36.14, 25]. The extension of this amplitude that includes the cosmological constant using a quantum deformation of the groups is defined in [36.1, 26].

There are three key properties of these amplitudes are three:

1. They define transition amplitudes for the Hilbert space $\mathcal{H}_\Gamma$, which have the correct degrees of freedom to describe the gravitational field.
2. They yield the Hamilton function of a truncation of Lorentzian general relativity over a cellular decomposition dual to C. More precisely, the vertex amplitude (36.9) yields the exponential of the Regge action on the corresponding cell [36.4] and the quantum deformed amplitude yields the exponential of the Regge action with cosmological content [36.1, 5]. The full amplitude is a truncation of a Feynman *sum over geometries* in the bulk [36.3].
3. They are finite. This is the key result. A theorem states that the vertex amplitude with cosmological constant is finite [36.1, 2]. The Planck length and the cosmological constant provide physical ultraviolet and infrared cutoff, respectively. (Without cosmological constant, they are still ultraviolet fi-

nite, but there are diverging radiative corrections describing large *spikes* of the geometry. These are cut off by the cosmological constant.)

Note the structure of the theory is shown in Fig. 36.4.

The amplitudes on a given truncation approximate the truncated dynamics of classical general relativity on a given triangulation. These approximate continuous general relativity when the truncation is refined. As in QCD and QED, the truncation is expected to offer already a good approximation to the full dynamics, in appropriate regimes.

Although a number of important technical issues, which must not be underestimated, remain open (see a discussion in [36.14]) Eqs. 36.8–36.9 give a definition of a quantum theory of gravity which is finite and yields classical general relativity in an appropriate limit. The construction provides a good ground for discussing the physical question of the nature of quantum spacetime.

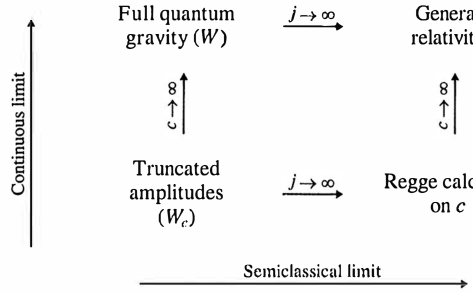


Fig. 36.4 The structure of covariant LQG

36.5 Quantum Spacetime

I can now summarize the conceptual structure that has emerged, for a quantum theory of spacetime.

Measurements directly involving the gravitational field are measurements of geometrical lengths, areas, or volumes. For instance, any measurement of a cross-section is the measurement of an area. Gravitational wave detectors measure (the variation of) a length. The outcomes of these measurements are described by the spin-network Hilbert space. No measurement measures an infinite number of quantities: we always have access only to a finite number of outcomes. Hence a truncation of the degrees of freedom is sufficient to describe the outcome of any measurement.

The theory predicts that measurements of some geometrical quantities yield discrete values [36.27, 28]. According to the theory, for instance, a physical cross section cannot be smaller than the Planck scale; it can only take values which are in the spectrum of the area operator. (See [36.29] for a discussion on the *diff invariance* of these prediction.)

Dynamics is given by associating an amplitude with each process. A process is determined by its boundary state, namely the outcome of a measurement (or a generic interaction) on its boundary. The relative probability of distinct processes can be computed from these amplitudes. The formalism does not require to go at infinite distance from an interaction to have well-defined physical amplitudes.

Spacial and temporal specifications make sense only on the boundary of a process, in the context of an interaction. In other words, space and time themselves are reduced to quantum entities such as the position of a quantum particle, which is determined only at interaction time, otherwise is fluctuating.

References

36.1 W.J. Fairbairn, C. Meusburger: Quantum deformation of two four-dimensional spin foam models, J.Math.Phys. **53**, 022501 (2012)
36.2 M. Han: 4-Dimensional spin-foam model with quantum Lorentz group, J. Math. Phys. **52**, 072501 (2011)
36.3 F. Conrady, L. Freidel: Path integral representation of spin foam models of 4-D gravity, Class. Quantum Gravity **25**, 245010 (2008)
36.4 J.W. Barrett, R.J. Dowdall, W.J. Fairbairn, F. Hellmann, R. Pereira: Lorentzian spin foam amplitudes: Graphical calculus and asymptotics, Class. Quantum Gravity **27**, 165009 (2010)
36.5 M. Han, M. Zhang: Asymptotics of spinfoam amplitude on simplicial manifold: Lorentzian theory, Class. Quantum Gravity **30**, 165012 (2013)
36.6 C. Rovelli: *Quantum Gravity* (Cambridge Univ. Press, Cambridge 2004)

36.7 C. Rovelli: Relational quantum mechanics, Int. J. Theor. Phys. **35**(9), 1637 (1996)
36.8 E. Bianchi: Talk at the 2012 Marcel Grossmann meeting (July 2012)
36.9 C. Rovelli: Discretizing parametrized systems: The magic of Ditt-invariance, arXiv:1107.2310 (2011)
36.10 A. Ashtekar: New variables for classical and quantum gravity, Phys. Rev. Lett. **57**, 2244 (1986)
36.11 E. Bianchi, P. Dona, S. Speziale: Polyhedra in loop quantum gravity, Phys. Rev. **D83**, 044035 (2011)
36.12 L. Freidel, S. Speziale: Twisted geometries: A geometric parametrisation of SU(2) phase space, Phys. Rev. **D82**, 084040 (2010)
36.13 M. Dupuis, J.P. Ryan, S. Speziale: Discrete gravity models and loop quantum gravity: A short review, Symmetry, Integrability and Geometry: Methods and Applications (SIGMA)8 (2012)
36.14 Rovelli: Zakopane lectures in loop gravity, Proc. 3rd Quantum Gravity Quantum Geometry School 003 (2011), arXiv:1102.3660
36.15 C. Rovelli, L. Smolin: Spin networks and quantum gravity, Phys. Rev. **D52**, 5743 (1995)
36.16 C. Rovelli, L. Smolin: Discreteness of area and volume in quantum gravity, Nucl. Phys. **B442**, 593 (1995)
36.17 A. Ashtekar, J. Lewandowski: Quantum theory of geometry. I: Area operators, Class. Quantum Gravity **14**, A55 (1997)
36.18 B. Bahr, F. Hellmann, W. Kaminski, M. Kisielowski, J. Lewandowski: Operator spin foam models, Class. Quantum Gravity **28**, 105003 (2011)
36.19 W. Wieland: Complex Ashtekar variables and reality conditions for Holst's action, Annales Henri Poincare **13**, 425 (2012)
36.20 J. Engle, E.R. Livine, R. Pereira, C. Rovelli: LQG vertex with finite Immirzi parameter, Nucl. Phys. **8799**, 136 (2008)
36.21 J. Engle, R. Pereira, C. Rovelli: The loop-quantum-gravity vertex-amplitude, Phys. Rev. Lett. **99**, 161301 (2007)
36.22 E.R. Livine, S. Speziale: A new spinfoam vertex for quantum gravity, Phys. Rev. **D76**, 084028 (2007)
36.23 L. Freidel, K. Krasnov: A new spin foam model for 4-D gravity, Class. Quantum Gravity **25**, 125018 (2008)
36.24 R. Pereira: Lorentzian LQG vertex amplitude, Class. Quantum Gravity **25**, 085013 (2008)
36.25 A. Perez: The spin foam approach to quantum gravity, Living Rev. Relativ. **16**, 3 (2013)
36.26 M. Han: Cosmological constant in LQG vertex amplitude, Phys. Rev. **D84**, 064010 (2011)
36.27 C. Rovelli: A Generally covariant quantum field theory and a prediction on quantum measurements of geometry, Nucl. Phys. **8405**, 797 (1993)
36.28 L. Smolin: Finite diffeomorphism invariant observables in quantum gravity, Phys. Rev. **D49**, 4028 (1994)
36.29 C. Rovelli: Comment on 'Are the spectra of geometrical operators in Loop Quantum Gravity really discrete?' by B. Dittrich and T. Thiemann, arXiv:0708.2481 (2007)

37. Gravity, Geometry, and the Quantum

Hanno Sahlmann

There are various indications that finding a quantum theory of gravity is important for a full understanding of fundamental physics. Loop quantum gravity is one possibility for such a quantum theory. In the following, we explain its origin in a gauge theoretic reformulation of gravity, and its status as a quantum theory of geometry. An overview is given over Einstein's equations in the quantum theory. As an example for an application of loop quantum gravity, the quantum theory of certain black hole horizons is sketched. We close with an outlook on current research and future challenges.

Perhaps the most surprising and delightful aspect about general relativity is that gravity is geometry. The gravitational field not only carries the gravitational interaction, it also determines the geometry of space–time. This dual nature of gravity makes contemplation of a quantum theory of gravity fascinating yet very difficult.

At the same time, a quantum theory of gravity carries with it the hope of resolving some of the problems that are posed by our present understanding of fundamental physics. First off, there is general relativity itself, in which singularities occur rather generically. Moreover, at these singularities, curvature diverges and space–time ends. We know that strong gravitational fields lead to particle creation via quantum processes, which in turn influences the gravitational field. Thus, these singularities are points at which one can not trust general relativity any more. Quantum theory and, ultimately, quantum gravity, must be taken into account, and it may lead to a drastic change in the picture, perhaps even by resolving the singularities.

Next, the quantum field theories (QFTs) that describe matter and its nongravitational interactions very well are often only effective theories. They contain terms that become very large at large energy scales or, differently put, at very small length scales. If quantum gravity changes the structure of space–time at the smallest scales, it has the potential to cut off these divergences, and result in a more fundamental picture for quantum field theory. Indeed, while there is no definite proof of nontrivial structure at small scales, there are some hints. With the Planck length

$$\ell_{\mathrm{P}} = \sqrt{\frac{G\hbar}{c^3}} \approx 1.6 \times 10^{-35}\,\mathrm{m}\,, \tag{37.1}$$

there exists a constant of nature that is a very small length scale. Natural scales often indicate that new phenomena are to be expected; thus, the Planck length may be indicative of a change of the structure of space–time at the smallest scales.

Last but not least, there are the mysterious laws of black-hole thermodynamics: stationary black holes are described by a few macroscopic parameters, just like a thermodynamical system in equilibrium. Moreover, the parameters obey equations that are mathemati-

cally equivalent to the equations of thermodynamics. This leads to the identification of thermodynamical and geometric quantities and, what is more, these identifications seem to make sense. For example, the black-hole mass is identified with thermodynamic energy, and black holes can also be assigned a temperature, due to their thermal Hawking emission. In this situation it is tempting to accept that black holes *are* thermodynamical systems, and to identify their microstates with states of the quantum theory of gravity.

How, then, can one approach the search for a quantum theory of gravity? In the absence of experimental hints, the only hard constraints are mathematical consistency and the reduction to general relativity in the appropriate circumstances. Therefore, the answers depend on the viewpoint.

One natural avenue to try is to quantize gravity perturbatively along the same lines as the other fundamental interactions, using a free graviton field on a fixed, typically flat, background and the apparatus of perturbation theory. This can give interesting results as long as one is only interested in an effective quantum theory describing physics in a limited regime, but if fails to give a fundamental theory due to the perturbative nonrenormalizability of gravity. One can add symmetries that improve the convergence, as has been explored in supergravity, or even new fundamental principles, as for example in string theory.

In contrast, loop quantum gravity (LQG) starts from the assumption that a perturbative approach and, in particular, a split into flat background and curvature perturbations does not match well the geometric character of gravity. Minkowski space, after all, has no special status among the solutions of Einstein's equations, besides the high degree of symmetry. Thus, loop quantum gravity is a nonperturbative approach to the quantum theory of gravity, in which no classical background metric is used. In particular, its starting point is not a linearized theory of gravity. Furthermore, loop quantum gravity starts from a classical formulation that is, on a kinematic level, identical to that of an SU(2) gauge theory [37.1, 2]. The starting point is thus much closer to that for the other fundamental interactions, which are also gauge theories.

Otherwise, loop quantum gravity is conservative, in the sense that new symmetries or additional gravitational or nongravitational fields are not essential. As we will explain, a certain amount of unification of the description of matter and gravity is achieved. However, the question of whether matter fields must have special properties to be consistently coupled to gravity in the

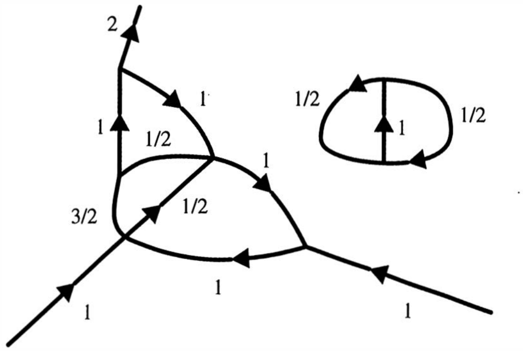


Fig. 37.1 A part of a spin network

framework of loop quantum gravity is one of the important open questions in loop quantum gravity. Finally, while loop quantum gravity still operates according to the rules of quantum field theory, the details are quite different from those for field theories that operate on a fixed classical background space–time.

One result that can be directly traced back to working without fixed background, and with gauge-theory variables, is that in loop quantum gravity, geometry is reduced to combinatorics and group representation theory. Indeed, aspects of loop quantum gravity were foreshadowed in works of *Penrose* (see for example [37.3]), who argued that geometry can emerge from the quantum theory of angular momentum. He studied the quantized geometry inherent in spin networks, directed graphs in which the edges carry representations of SU(2) (labeled by spin quantum numbers), and the vertices are invariant tensors under the action of SU(2) given by the representations going into, or out of, the vertex (Fig. 37.1).

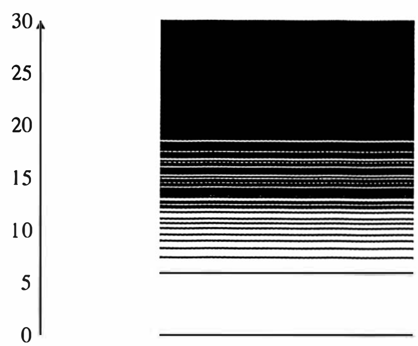


Fig. 37.2 The lowest part of the area spectrum of loop quantum gravity, in units of l_P^2

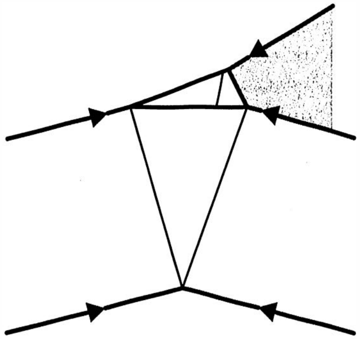


Fig. 37.3 Part of a spin-foam diagram

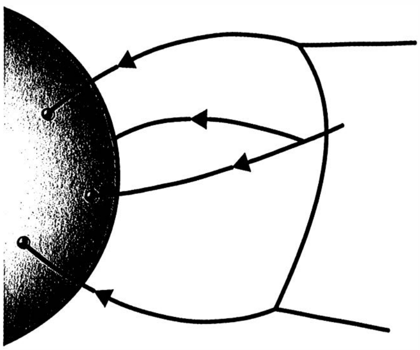


Fig. 37.4 Sketch of a black hole in loop quantum gravity: spin networks end on the horizon, endow it with area, and couple to the surface degrees of freedom

While Penrose was going for mathematical simplicity and beauty, with the development of loop quantum gravity his conjectures became the result of a derivation: in loop quantum gravity spin networks are a representation for quantum states of spatial geometry [37.4–8]. In this quantum theory of geometry, which will be the topic of Sect. 37.2, quantities such as area and volume are quantized in units of the order of the Planck length. In this sense, there is indeed a minimal length scale of the order of the Planck length in loop quantum gravity. But, the spectrum becomes exponentially dense for large eigenvalues, thus effectively forming a continuum. As an example, the lowest part of the area spectrum is depicted in Fig. 37.2.

The edges of the spin networks can be thought of as flux tubes of area, whereas the vertices carry information about volumes.

Loop quantum gravity uses, in its original version, a canonical approach to quantization. Einstein's equations are encoded into constraints on the phase-space variables. These translate into operator equations in the quantum theory. The spin-network states do not satisfy the quantum version of Einstein's equations. Rather, extremely complicated linear combinations have to be formed. They can be organized into spin-foam diagrams, which can be considered as depicting a kind of time evolution of spin networks [37.9]. A part of such a diagram is depicted in Fig. 37.3 and a detailed discussion can be found in Chap. 38.

While the interpretation of these as diagrams as quantized space–times is not yet understood in all details, it is very encouraging that one can arrive at the same structure from a path-integral point of view: in spin-foam gravity, these diagrams describe an expansion of a discretized path integral for gravity. A separate chapter of this handbook is devoted to this approach to quantum gravity. Recent developments show that spin-foam gravity can literally be regarded as a path-integral formulation of LQG [37.10–16].

It is also very encouraging that loop quantum gravity provides a description of black-hole horizons that can account for black-hole entropy for static and rotating, charged, and neutral black holes. Spin networks can end on the horizon, endow it with area, and couple to the surface degrees of freedom, which are described by a Chern–Simons gauge theory; see Fig. 37.4.

The space of microstates of the horizon accounts for the entropy.

Another area of success for loop quantum gravity is cosmology. In loop quantum cosmology, one applies the principles of loop quantum gravity to minisuperspace models for cosmology. It turns out that in this context, the big bang singularity is indeed resolved, and inflation is favored [37.17–21]. This area is discussed in detail in Chap. 39.

But, there are also a number of challenges. Some of these arise from the fact that the requirement of background independence leads to a theory which is built around a very quantum mechanical gravitational vacuum, a state with degenerate and highly fluctuating geometry. On the one hand, this is exciting, because it means that when working in loop quantum gravity, the deep quantum regime of gravity is *at one's fingertips*. However, it also means that to make contact with low-energy physics is a complicated endeavor. The latter problem has attracted a considerable amount of work, but is still not completely solved. A better understanding of this issue should also lead to unequivocal testable predictions from loop quantum gravity, which are so far missing.

A related area of work concerns the generation and interpretation of solutions: in loop quantum gravity the question of finding quantum states that satisfy quantum Einstein equations is reformulated as finding states that are annihilated by the quantum Hamilton constraint. The choices that go into the definition of this constraint are not yet well understood in physical terms. In particular, the constraint should be implemented in an anomaly-free way, but what this entails in practice, and whether existing proposals fulfill this requirement, are still under debate. Also, methods for approximately solving the constraint, and the interpretation of the solutions, are still under investigation.

Loop quantum gravity is a large research area, and the present chapter can not replace some more complete and detailed reviews that are available. We refer the interested reader in particular to [37.22–24], and to the textbook [37.25].

37.1 Gravity as a Gauge Theory

In the present section, we will briefly describe the classical phase-space description that underlies loop quantum gravity.

The original phase-space description of gravity is due to *Arnowitt*, *Deser*, and *Misner* (ADM) [37.26], described in Chaps. 16–18. The phase-space variables are the spatial metric q_{ab} of a spatial slice of space–time, and $\pi^{ab} = \sqrt{q}(K^{ab} - Kq^{ab})$ with K^{ab} the extrinsic curvature of the slice. They are canonical coordinates, with the Poisson brackets

$$\left\{q_{ab}(p), K^{a'b'}(p')\right\} = 8\pi G\delta_a^{a'}\delta_b^{b'}\delta(p,p') \,. \tag{37.2}$$

But these coordinates are not free. Rather, they have to satisfy two sets of constraints

$$\begin{aligned} D_a\pi^{ab} &= 0\,, \\ \sqrt{q}\left(R[q] - \frac{1}{q}\left(\frac{\pi^2}{2} - \pi^{ab}\pi_{ab}\right)\right) &= 0\,. \end{aligned} \tag{37.3}$$

In the absence of nontrivial boundary conditions, the canonical Hamiltonian is just a combination of these constraints. So, evolution under the Hamiltonian is gauge, and has to be divided out to obtain the true phase space.

A starting point for loop quantum gravity is now the following remarkable fact [37.1, 2, 27, 28]:

> *There is a formulation of general relativity in which the (unconstrained) phase space is precisely that of SU(2) Yang–Mills theory.*

This phase-space description is a natural basis for a quantization of gravity. The other fundamental interactions are described by Yang–Mills theories; thus, one obtains a more unified description of all interactions. Moreover, this description makes the tools available that have been developed for quantizing the other interactions. The original approach was an extension of the phase space (37.2) and (37.3), followed by a canonical transformation [37.1, 2]. In the following, we will derive the canonical formulation from a covariant action principle. For some more detail on this, see [37.22].

A first step toward a formulation of gravity as a gauge theory is the introduction of the connection as an independent degree of freedom. This can be done in the standard Einstein–Hilbert action, but more interesting for us is the first-order formalism. Here, the Palatini action reads

$$S_{\mathrm{P}}[\omega, e] = \frac{1}{4\kappa}\int \mathrm{d}^4x \epsilon_{IJKL} e^I \wedge e^J \wedge F(\omega)^{KL} \,, \tag{37.4}$$

where $\kappa = 8\pi G$ is the coupling constant of gravity. The field e is a coframe, that is, a pointwise basis of the cotangential bundle $e^I \equiv e^I_\mu \,\mathrm{d}x^\mu$. Equivalently, it can be viewed as a point-dependent map $\mathbb{R}^4 \to T^pM$. It defines a space–time metric via

$$g_{\mu\nu} = e^I_\mu e^J_\nu \eta_{IJ} \,. \tag{37.5}$$

Vice versa, a space–time metric g defines an orthonormal coframe via (37.5), but only up to SO(3, 1) rotations in the internal space $\mathbb{R}^4$. ω on the other hand is an SO(3,1) connection, and F its curvature.

A Legendre transform of the Palatini action leads (after solving the second-class constraints) back to the original ADM variables (37.2). But, this can be changed by adding the Holst term to the action

$$S[\omega, e] = S_{\mathrm{P}}[\omega, e] + S_{\mathrm{H}}[\omega, e] \,, \tag{37.6}$$

with [37.29]

$$S_{\mathrm{H}}[\omega, e] = -\frac{1}{2\kappa\beta}\int \mathrm{d}^4x e^I \wedge e^J \wedge F(\omega)_{IJ} \,. \tag{37.7}$$

The new term comes with an additional coupling constant β, the Barbero–Immirzi parameter. The equations $\delta S/\delta\omega$ are equivalent to $D^{(\omega)}e = 0$ [37.29], which can be solved for $\omega \equiv \omega(e)$. Re-inserting $\omega(e)$ into the action gives

$$S_{\rm H}[\omega(e), e] = 0\,, \quad S_{\rm P}[\omega(e), e] = S_{\rm EH}[g(e)]\,, \tag{37.8}$$

where $S_{\rm EH}$ is the Einstein–Hilbert action. Thus, the above action principle leads to the equations of motion of general relativity, irrespective of the value of β. It *does* have physical significance in the case that spinor matter is coupled to gravity [37.30]. Solving $\delta S/\delta\omega = 0$ and re-inserting into the action gives an effective four-fermion interaction in that case with β-dependent strength. The effect is however suppressed by the gravitational coupling constant, and is thus extremely small.

For the values $\beta = \pm\mathrm{i}$, the resulting canonical formulation has special properties [37.1, 2], which will be briefly described below.

To go over to the Hamiltonian formulation, one chooses a time function t on the space–time manifold, whose level surfaces Σ_t give a foliation of the space–time into spatial slices, as well as a time covector field t^α with $t^\alpha_\alpha\partial t = 1$, and decompose it into tangential and normal components with respect to Σ_t

$$t^\alpha = Nn^\alpha + N^\alpha\,, \tag{37.9}$$

where n^α is the unit normal and the shift vector N^α is tangential. The gauge freedom in the coframe is partially fixed by going to time gauge

$$e^0_\mu = n_\mu\,. \tag{37.10}$$

Since n_μ is time like, only SO(3) remains as gauge group. The covariant fields can now be decomposed accordingly, and adapted coordinates be chosen. The coframe assumes the structure

$$(e^I_\mu) = \begin{pmatrix} N & N^i \\ 0 & \\ 0 & e^i_a \\ 0 & \end{pmatrix}\,, \tag{37.11}$$

where we let I run horizontally and μ vertically. The index i now runs from one to three, and a is a tangent-space index for Σ_t. Analogously, ω can be decomposed into SO(3) connections $\Gamma^i_a := \epsilon^{i0}{}_{KL}\omega^{KL}_a$, $K^i_a := \omega^{i0}_a$ and the rest, i. e., the components of ω_0. The action can now be expressed in terms of the decomposed fields [37.22]

$$S = \frac{1}{\kappa\beta}\int \mathrm{d}t \int_{\Sigma_t} E^a_i \dot{A}^i_a - \underbrace{\left(\omega^i_0 G_i + N^a C'_a + NC'\right)}_{=:h_{\rm can}}\,, \tag{37.12}$$

with

$$\begin{aligned} A^i_a &= \Gamma^i_a + \beta K^i_a\,, \\ E^a_i &= \sqrt{\det q} e^a_i\,, \\ q_{ab} &= e^i_a e^j_b \delta_{ij}\,. \end{aligned} \tag{37.13}$$

Here, q denotes the metric on Σ_t and the dot is the time derivative $t^\alpha_\alpha\partial$. We see that A and E are conjugate canonical variables

$$\{A^i_a(x), E^b_j(y)\} = \kappa\beta\delta^b_a\delta^i_j\delta(x,y)\,. \tag{37.14}$$

These variables were first introduced in [37.1, 2] for the special case $\beta = \pm\,\mathrm{i}$ by a canonical transformation from, and extension of, the phase-space formulation in terms of ADM variables.

The Hamiltonian density $h_{\rm can}$ is a sum of constraints. One has

$$\begin{aligned} G_i &= D^{(A)}_a E^a_i\,, \quad C_a = E^b_i F^i_{ab}\,, \\ C &= \frac{\beta}{2}\frac{E^a_i E^b_j}{\det E}\left[\epsilon^{ij}{}_k F^k_{ab} - 2(1+\beta^2)K^i_{[a}K^j_{b]}\right]\,, \end{aligned} \tag{37.15}$$

where multiples of G_i were subtracted from G'_a and C' to obtain G_a and C. The constraint equations $G_i = G_a = C = 0$, together with the evolution equations

$$\{A(x), h_{\rm can}\} = \dot{A}(x)\,, \quad \{E(x), h_{\rm can}\} = \dot{E}(x)\,, \tag{37.16}$$

are completely equivalent to Einstein's equations. But, as is the case for all reparameterization-invariant systems, time evolution is a gauge transformation. Concretely

$$\begin{aligned} \{A_a, G(\Lambda)\} &= -D^{(A)}_a \Lambda\,, \\ \{E^a, G(\Lambda)\} &= [\Lambda, E^a]\,, \end{aligned}$$

so G generates gauge transformations on the unreduced phase space. Moreover

$$\{A, \boldsymbol{C}(N)\} = \mathcal{L}_N A\,, \quad \{E, \boldsymbol{C}(N)\} = \mathcal{L}_N E\,, \tag{37.17}$$

so $\boldsymbol{C}$ generates diffeomorphisms of the spatial slice Σ. C is related to $\mathcal{L}_{Nn^\alpha}$, i.e., it generates the diffeomorphisms in a direction normal to Σ, as long as all the fields are on-shell. The constraints form an algebra, the Dirac algebra

$$\begin{aligned}\{G(\Lambda), G(\Lambda')\} &= G([\Lambda', \Lambda])\,,\\ \{G(\Lambda), \boldsymbol{C}(N)\} &= -G(\mathcal{L}_N \Lambda)\,,\\ \{\boldsymbol{C}(N), \boldsymbol{C}(N')\} &= \boldsymbol{C}([N, N'])\,.\end{aligned} \tag{37.18}$$

The Hamiltonian constraint C is gauge invariant and transforms under diffeomorphisms in the expected way

$$\begin{aligned}\{C(N), G(\lambda)\} &= 0\,,\\ \{C(N), \boldsymbol{C}(N)\} &= C(\mathcal{L}_N N)\,.\end{aligned} \tag{37.19}$$

Up to here, the structure is that of an infinite-dimensional Lie algebra. But, the bracket of two Hamiltonian constraints is more complicated

$$\begin{aligned}\{C(N), C(M)\} &= -\frac{\kappa^2\beta^2}{4}\boldsymbol{C}(S)\,,\\ \text{with } S^a &= \frac{E^a E^b}{\det q}(N\partial_b M - M\partial_b N)\,.\end{aligned} \tag{37.20}$$

It contains a function of the phase-space point on the right-hand side, hence the structure is not that of a Lie algebra any more.

The constraints commute on the constraint surface $G = \boldsymbol{C} = C = 0$. This means that they form a *first-class system*, and thus the constraints can be imposed in the quantum theory as operator equations. This is the result of imposing the time gauge. Without it, the situation is more complicated, but also very interesting [37.31–39].

Some remarks are in order:

1. It is instructive to compare the canonical formulation given above to that of electrodynamics. In that case, the action can be written

$$\begin{aligned}S[A] &= -\tfrac{1}{4}\int_{\mathbb{M}} F_{\mu\nu}F^{\mu\nu}\,\mathrm{d}^4x\\ &= \int \mathrm{d}t \int \mathrm{d}^3x - E^a\dot{A}_a - \frac{1}{2}(E^2 + B^2) + A_0\nabla\cdot E\,,\end{aligned} \tag{37.21}$$

with A the 4-potential, F its curvature, and Σ_t an equal-time surface relative to some inertial observer time t. One recognizes the vector potential A and the electric field E as the analogues of the gravitational phase-space variables. Also, the constraint $\nabla\cdot E$ is analogous, as it generates the U(1) gauge transformations. But, the rest of the structure is different. There are no diffeomorphism and Hamiltonian constraints, due to the fact that the metric is fixed in (37.21). Instead, there is a nonvanishing canonical Hamiltonian generating physical time evolution.
2. The above provided a canonical formulation of Einstein gravity in $D+1 = 4$ dimensions in terms of a phase space that is identical to that of SO(D) Yang–Mills theory. This formalism relies on a coincidence that only happens for $D = 3$: an SO(D) connection has $D(D-1)/2$ components, whereas a spatial frame has D. If they are to be canonically conjugate variables, they have to have the same number of components, which restricts to $D = 3$. There are, however, similar formulations for higher dimensions, with additional constraints [37.40–43], and also 2 + 1 gravity can be formulated as a gauge theory [37.44].
3. In the following, we will go over from a formulation in terms of SO(3) to one in terms of its covering group SU(2). This must be done in order to couple fermions to gravity, but it does not change classical or quantum theory much. The biggest change is that representations with half-integer spin will also be allowed in the quantum theory.
4. Matter fields can be added to the canonical description given above. This has to be done with some care, so as not to change the structure of the gravitational sector. For the fermionic sector this requires working with slightly unusual (*half-density*) variables [37.45].

37.2 Quantum Geometry

37.2.1 Kinematic Quantization

We will now discuss the quantization of the unreduced phase space of gravity, coordinatized by the fields A and E of (37.13). This is the first step in the algorithm for the quantization of constrained systems devised by Dirac (for the original account, see his *Lectures on Quantum Mechanics*, for a modern treatment, see [37.23]):

1. Quantization of the unreduced phase space (*kinematic quantization*)
2. Imposition of the constraints as operator equations on states. The solutions of these equations are the physical quantum states of the system.

The kinematic quantization will be discussed in the following two sections.

The functions in the unreduced phase space have a geometrical interpretation via intrinsic and extrinsic curvatures of a spatial slice of space–time. Thus, states in the kinematic Hilbert space have an interpretation in terms of quantized geometry. This quantized geometry has been studied extensively in loop quantum gravity. It is the subject of Sect. 37.2.4.

37.2.2 The Holonomy-Flux Algebra

To quantize the unreduced phase space, one is seeking a representation of the canonical commutation relations

$$\left[A_a^i(x), E_j^b(y)\right] = \kappa\hbar\beta\delta_a^b\delta_j^i\delta(x,y) \tag{37.22}$$

on some Hilbert space. Note that $\kappa\hbar \propto \ell_P^2$, the *Planck area.* Fields evaluated at points are usually too singular to give well-defined operators in the quantum theory. Thus, one has to form suitably integrated quantities that correspond to well-defined operators in the quantum theory. Poisson brackets then suggest commutation relations for these quantities, and one defines an abstract algebra of operators.

Details matter in this context, as different choices of smearing can lead to different algebras and hence to different quantum theories. In loop quantum gravity a different choice of algebra is made than is customary in Yang–Mills theory [37.4, 46, 47]. In the latter case, both the algebra and its representations used in the quantum theory make use of the metric as a classical background field in their construction. In general relativity, the metric is dynamic and hence can not be used as a background field. Moreover, a splitting of the metric into background and dynamical parts, while very useful in practical applications, is not natural from a fundamental perspective.

Hence, the algebra and its representation chosen in loop quantum gravity do not make use of any background metric. To illustrate the different choices, it is instructive to first consider the case of electromagnetism. The usual quantum theory is obtained by declaring

$$\begin{aligned}&\left[\int f^a A_a \sqrt{\det q}\,\mathrm{d}^3x, \int f_b' E^b\,\mathrm{d}^3x\right]\\&= \mathrm{i}\hbar \int f^a f_a' \sqrt{\det q}\,\mathrm{d}^3x\,\mathrm{id}\,,\end{aligned} \tag{37.23}$$

and by defining the ground state as a Fock state, by

$$a(f)\Omega = 0\,. \tag{37.24}$$

Here, f, f' are arbitrary smearing functions, and the definition of the annihilation operators a makes use of the metric q in various ways. But, one could also define

$$E(S) := \int\limits_S E^a \epsilon_{abc}\mathrm{d}x^b \wedge \mathrm{d}x^c\,, \tag{37.25}$$

where S is an oriented surface and ϵ_{abc} is the tensor density that is equal to the totally antisymmetric symbol in any coordinate system. We note that E^a has density weight $+1$ whereas ϵ_{abc} has weight -1, so the integrand is a two-form and the integral, using the orientation of S, is hence coordinate independent. Similarly, defining

$$A(e) = \int\limits_e A_a \mathrm{d}x^a\,, \tag{37.26}$$

where e is a curve, one finds (by a limiting procedure from the Poisson brackets of the point fields)

$$[A(e), E(S)] = \mathrm{i}\hbar I(e,S)\mathrm{id}\,, \tag{37.27}$$

where $I(e,S)$ is the signed intersection number for S and e. $I(e,S)$ is a purely topological quantity. The metric has thus dropped out of all definitions and relations.

A similar thing can be done for gravity. For a surface S and a Lie algebra valued smearing field n on S, one defines

$$E_n(S) := \int\limits_S n^i E_i^a \epsilon_{abc}\mathrm{d}x^b \wedge \mathrm{d}x^c\,. \tag{37.28}$$

For A, we use the quantity analogous to $\exp(iA(e))$. We choose a local trivialization and define the *holonomy*

$$h_e := \mathcal{P}\exp\int_e A \tag{37.29}$$

$$= \mathbb{I} + \int_0^1 A(e(t))\dot{e}^a(t)\mathrm{d}t + \int_0^1 \mathrm{d}t_1 \int_{t_1}^1 \mathrm{d}t_2 A_a(t_1)\dot{e}^a(t_1)A_a(t_2)\dot{e}^a(t_2) + \ldots\,, \tag{37.30}$$

which is an element of SU(2), and gives the parallel transport map from the fiber over the beginning point $b(e)$ of the edge to the fiber over of its final point $f(e)$, for the chosen trivialization. The reason for the choice of holonomies is that A, as a one-form, can be integrated over a one-dimensional oriented submanifold without use of a metric. In comparison to the simpler variables (37.26), holonomies offer the additional benefit that they transform in a simple way under gauge transformations, i. e., changes of trivialization, $g: M \to SU(2)$

$$h_e \mapsto g(b(e))h_e g(f(e))^{-1}\,. \tag{37.31}$$

In particular, traces of holonomies around closed loops are invariant.

The commutator is slightly more complicated than in the electromagnetic case

$$[E_n(S), h_e] = \begin{cases} \beta\kappa h_{e_1}\tau_j n^j(p) h_{e_2} & \text{if } S \cap e = \{p\}\,, \\ 0 & \text{if } S \cap e = \emptyset\,, \end{cases} \tag{37.32}$$

where $\{\tau_j\}$ is a basis of SU(2), and in the first line it was assumed that there is a single transversal intersection p between S and e. If the intersection is tangential, the commutator vanishes as well. Again, one sees that the commutator just depends on relative properties of S and e that are invariant under diffeomorphisms.

It is convenient to slightly generalize these variables. Given a *graph* of paths $\gamma = \{e_1, e_2, \ldots, e_n\}$ and a function $f: \mathrm{SU}(2)^n \to \mathbb{C}$, one obtains the functional

$$f_\gamma[A] := f(h_{e_1}[A], h_{e_2}[A], \ldots, h_{e_n}[A])\,. \tag{37.33}$$

A functional f is called *cylindrical with respect to* γ (written $f \in \mathrm{Cyl}_\gamma$) if it is of the above form, and simply *cylindrical* if it is of the above form for *some* graph γ. We note that

- A given cylindrical functional is cylindrical on *many* graphs. Consider the example of a function $f_\gamma[A] = f(h_e[A])$, which is cylindrical w.r.t. the graph $\gamma = \{e\}$. Now consider a second graph $\gamma' = e_1, e_2, e_3$, with $e_1 \circ e_2 = e$, and e_3 independent of e. Then f_γ is also cylindrical w.r.t. γ', as it can be written purely in terms of holonomies along edges in γ', $f_\gamma[A] = f(h_{e_1}[A]h_{e_2}[A])$.
- For two cylindrical functions which are cylindrical on graphs with *smooth* edges, one can not always find a *finite* graph such that they are both cylindrical w.r.t. to that graph, because they can intersect infinitely many times. But, for more regular edges, for example analytic or semi-analytic (roughly speaking: piecewise real analytic [37.48]) ones this can not happen, and so one can always find such a graph. As a consequence, such cylindrical functions are closed under addition and multiplication and thus form an algebra, denoted Cyl. In the following, we will always assume edges (and also surfaces) to be semi-analytic.

One can use these observations to write the commutator between the canonical variables in a relatively concise form. Because of the observations made above, one can assume without loss of generality that a surface S and a graph γ intersect only in vertices of γ. The commutator then reads

$$[f_\gamma, E_n(S)] \equiv X_n(S)(f_\gamma) = \frac{\beta\kappa}{2}\sum_{v\in V(\gamma)} n^i(v) \times \left[\sum_{e \text{ at } v} \kappa(S,e,v)\widehat{J}_i^{(v,e)} f\right](h_{e_1}, h_{e_2}, \ldots)\,, \tag{37.34}$$

where $V(\gamma)$ denotes the set of all vertices of γ

$$\kappa(S,e,v) = \begin{cases} 0 & \text{if } e \text{ intersects } S \text{ tangentially in } v \text{ or does not intersect } S \text{ at all} \\ 1 & \text{if } e \text{ intersects } S \text{ transversally and is above } S \\ -1 & \text{if } e \text{ intersects } S \text{ transversally and is below } S \end{cases} \tag{37.35}$$

and

$$\widehat{J}_k^{(v,e)} = \text{id} \otimes \text{id} \otimes \cdots \otimes \left\{ \begin{array}{c} L_k^e \\ R_k^e \end{array} \right\} \otimes \text{id} \otimes \ldots ,$$
$$\text{when } \left\{ \begin{array}{c} e \text{ ingoing at } v \\ e \text{ outgoing at } v \end{array} \right\} . \qquad (37.36)$$

Here, R and L denote the right/left invariant vector fields on SU(2) associated with a basis τ_k of su(2). The additional factor of 1/2 in (37.34) as compared to (37.32) is due to the assumption that edges must end on the surface. An edge that continues on both sides of the surface as in (37.32) will count as two separate edges in (37.34).

For general surfaces, the commutator above may not be a cylindrical function again, because edges and surfaces can intersect each other infinitely often. Thus, one must also restrict the surfaces to be in a suitable regularity class, such as semi-analytic or real analytic. Then the operation $X_n(S)$ defined above is a derivation on the space Cyl of cylindrical functions, i. e., it satisfies the Leibniz property

$$X_n(S)(fg) = X_n(S)(f)g + fX_n(S)(g) . \qquad (37.37)$$

Also, the commutator must have the Jacobi property, so

$$[f, [E_n(S), E_{n'}(S')]] = [X_n(S), X_{n'}(S')](f) \qquad (37.38)$$

and the commutator on the right-hand side is *nonvanishing* in general. Thus, we find that the operators corresponding to the spatial geometry do not commute.

The objects $E_n(S)$, together with the cylindrical functions Cyl subject to the above commutator relations, form the holonomy-flux algebra $\mathfrak{A}$. Since it does not make reference to classical geometry on Σ, diffeomorphisms ϕ acting in a simple way

$$\alpha_\phi(f)[A] := f(\phi_* A) , \quad \alpha_\phi(E_n(S)) = E_{\phi_* n}(\phi(S)) \qquad (37.39)$$

are automorphisms of $\mathfrak{A}$. A similar statement holds for gauge transformations.

37.2.3 The Ashtekar–Lewandowski Representation

To implement the constraints, one has to find a representation of the holonomy-flux algebra $\mathfrak{A}$, i. e., a mapping of $\mathfrak{A}$ into the operators of a Hilbert space that preserves the algebra structure. There are many representations of $\mathfrak{A}$, but again the nature of gravity can be a guide. Since there is no preferred metric in general relativity, one is looking for a representation that does not single out a preferred geometry. This is the case for the *Ashtekar–Lewandowski* representation of $\mathfrak{A}$ [37.7]. In fact, one can show that it is the only such representation, in a precise technical sense [37.48, 49].

Roughly speaking, the states in this representation are generated from a vacuum state that is invariant under diffeomorphisms, by the action of the holonomy operators. The operators $E_n(S)$ act via the derivations $X_n(s)$ of (37.34).

To construct this representation, note first that an inner product on Cyl can be defined by

$$\langle f_\gamma | f'_\gamma \rangle := \int_{\text{SU}(2)^n} \text{d}\mu(g_1) \ldots \text{d}\mu(g_n) \times \overline{f(g_1, g_2, \ldots, g_n)} f'(g_1, g_2, \ldots, g_n) . \qquad (37.40)$$

The measure $\text{d}\mu$ used above is the Haar measure on SU(2), and we have assumed without loss of generality that the two functions are cylindrical w.r.t. the same graph, as discussed below (37.33). Closure with respect to the corresponding norm gives a Hilbert space $\mathcal{H}_{\text{kin}}$. It can be shown that this space has a very suggestive structure, $\mathcal{H}_{\text{kin}} = L^2(\overline{\mathcal{A}}, \text{d}\mu_{\text{AL}})$, the square-integrable functions over a space of distributional connections, with respect to a certain measure [37.50], which can be interpreted as a kind of Lebesgue measure on the space of connections.

The action of the basic operators in this representation is analogous to that found in the Schrödinger representation of quantum mechanics

$$\pi(f)\Psi[A] = f[A]\Psi[A] ,$$
$$\pi(E_n(S))\Psi[A] = (X_n(S)\Psi)[A] , \qquad (37.41)$$

where we have assumed that Ψ is smooth enough for $X_n(S)$ to act. For example, Ψ could be a cylindrical function based on a differentiable function on some power of SU(2). But, the properties of this representation are very different from those of the Schrödinger representation of quantum mechanics, and of the representations encountered in standard QFT. For example, eigenstates of the fluxes, i. e., the momentum variables, are normalizable, as we will see in a moment. Also, there are precise analogues of this representation for

scalar and gauge fields, and they are unitarily inequivalent to the standard representations for those fields.

The representation has several useful properties: it is irreducible and faithful. No background geometry was used in the definitions, so it carries a unitary representation of spatial diffeomorphisms and gauge transformations. It has an invariant vacuum, from which all states can be generated by the action of the basic operators.

The Hilbert space $\mathcal{H}_{\text{kin}}$ has a very useful orthonormal basis. It is precisely labeled by the spin networks that were mentioned in the introduction, and some slight extensions of them. To explain, let us first consider a general compact Lie group G. Then there are two natural representations of G on $\mathcal{H}_G = L^2(G, \mathrm{d}\mu)$, the left- and right-regular representations

$$(\rho_{\mathrm{L}}(g)f)(g') = f(g'g^{-1}) , \quad (\rho_{\mathrm{R}}(g)f)(g') = f(gg') . \tag{37.42}$$

They both decompose into irreducible representations (irreps) and, since the two representations commute, there is a common basis of eigenvectors of the Casimir operators. Let π be an irrep of G; then

$$V(\pi, m) := \mathrm{span}\{\pi_{mn}(\cdot) | n = 1, 2, \ldots, \dim \pi\}$$
$$\text{is left invariant by } \rho_{\mathrm{L}} , \tag{37.43}$$

$$\overline{V}(\pi, n) := \mathrm{span}\{\pi_{mn}(\cdot) | m = 1, 2, \ldots, \dim \pi\}$$
$$\text{is left invariant by } \rho_{\mathrm{R}} . \tag{37.44}$$

The representation $V(\pi,m)$ induced by ρ_{L} is π itself. The one induced by ρ_{R} on $\overline{V}(\pi, n)$ is its dual, $\overline{\pi}$, i. e., $\overline{\pi}(g) = \pi(g^{-1})^T$. The Peter–Weyl theorem now asserts that each irrep arises in the decomposition of the regular representations and, even more, that their matrix elements give a basis of $\mathcal{H}_G$. Pick, for each equivalence class of irreps of G, a representative π; then the set of all $\sqrt{\dim \pi}\pi_{mn}$ for all equivalence classes forms an orthonormal basis of $\mathcal{H}_G$.

Let $\mathcal{H}_\gamma = \overline{\mathrm{Cyl}_\gamma}^{\|\cdot\|}$. On the one hand, $\mathcal{H}_\gamma$ is a subspace of $\mathcal{H}_{\text{kin}}$; on the other hand, it is isomorphic to $L^2(\mathrm{SU}(2)^n)$. Thus, an orthonormal basis for $\mathcal{H}_\gamma$ is given by

$$\left(\prod_i (2j_i + 1)\right)^{1/2} \prod_i \pi^{j_i}_{k_i l_i}(h_{e_i}[A]) , \tag{37.45}$$

where the $j_1, j_2, \ldots, j_n$ label irreducible representations of SU(2). The only problem with this decomposition is that in general $\mathcal{H}_\gamma$ is not orthogonal to $\mathcal{H}_{\gamma'}$ for graphs γ, γ' that overlap. Take for example $\gamma = \{e\}$, $\gamma' = \{e_1, e_2\}$ with $e = e_1 \circ e_2$. Then

$$\pi_{mn}(h_e[A]) = \sum_{m'} \pi_{mm'}(h_{e_1}[A]) \pi_{m'n}(h_{e_2}[A]) . \tag{37.46}$$

Therefore, one introduces a family of slightly modified Hilbert spaces $\mathcal{H}'_\gamma$, which give a decomposition of $\mathcal{H}_{\text{kin}}$ into orthogonal subspaces. To describe it, we need to discuss the transformation properties of vectors under gauge transformations.

We start by considering just a single edge e. With respect to gauge transformations g, the vectors $\pi^j_{mn}(h_e)$ transform under the tensor product $j \otimes \bar{j}$, and can be visualized as the edge with representation j sitting at its end point and representation $\bar{j}$ at its starting point. When several edges meet at a vertex v, contractions of the matrix indices of the representation matrices at that vertex can be done and correspond to vectors in the tensor product

$$\mathcal{H}_v = \left(\bigotimes_{e \text{ into } v} j_e\right) \otimes \left(\bigotimes_{e \text{ out of } v} \bar{j}_e\right) . \tag{37.47}$$

To give an orthogonal basis of this space, one can simply decompose it into irreps

$$\mathcal{H}_v = \bigoplus_l c_l l , \tag{37.48}$$

where c_l counts the multiplicity of the spin l-representation. When we apply this to the situation in LQG, we obtain the following decomposition. Given a graph γ

$$\mathcal{H}_\gamma = \bigoplus_j \mathcal{H}_{\gamma,j} = \bigoplus_{j,l} \mathcal{H}_{\gamma,j,l} . \tag{37.49}$$

Here, we have first decomposed into spaces in which the assignment of irreps to edges (labeled by $\boldsymbol{j}$) is fixed, giving essentially the tensor products of the spaces (37.47). Then we have further decomposed according to (37.48), labeling the irreducible subspace chosen at the vertices with $\boldsymbol{l}$.

Now one can remedy the problem that the decomposition into $\mathcal{H}_\gamma$ was not an orthogonal one. Given again

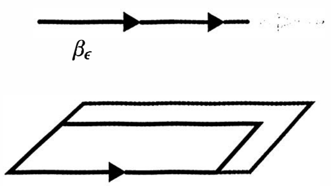


Fig. 37.5 The loop α_ϵ and the edge β_ϵ of (37.53)

a graph γ, we can call a labeling $\boldsymbol{j}$ of edges and $\boldsymbol{l}$ of vertices with irreps *admissible* if no two-valent vertex has been assigned the trivial representation $l = 0$, and none of the irreps assigned to the edges is trivial. Then we set

$$\mathcal{H}'_\gamma = \bigoplus_{\boldsymbol{j},\boldsymbol{l}\text{ admissible}} \mathcal{H}_{\gamma,\boldsymbol{j},\boldsymbol{l}} \tag{37.50}$$

and obtain the desired orthogonal decomposition

$$\mathcal{H} = \bigoplus_\gamma \mathcal{H}'_\gamma \,. \tag{37.51}$$

We finish this section by noting that the AL-representation has the following peculiar properties:

1. Diffeomorphisms ϕ are represented on $\mathcal{H}_{\text{kin}}$ by unitary operators U_ϕ. But, generators for these unitary operators do not exist. If $\phi(t)$ is a one-parameter family of diffeomorphisms, with $\phi(0) = \mathbb{I}$, then

$$\frac{1}{\mathrm{i}} \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_0 U_{\phi(t)} \tag{37.52}$$

   does not exist, in any sense, as a well-defined operator.
2. We have seen that the holonomies $h_e[A]$ exist as matrices of operators. But, can one obtain from them an operator neither for the curvature F nor for the connection A itself: the limits

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon^2} (h_{\alpha_\epsilon} - \mathbb{I}) \,, \quad \lim_{\epsilon\to 0} \frac{1}{\epsilon} \left(h_{\beta_\epsilon} - \mathbb{I}\right) \tag{37.53}$$

   do not exist in any sense as well-defined operators on $\mathcal{H}_{\text{kin}}$. α_ϵ is here a plaquette loop with (coordinate) side length ϵ, and β_ϵ is an open line with (coordinate) side length ϵ (Fig. 37.5).

37.2.4 Geometric Operators

Now we return to the geometric nature of the fields and explore the geometry residing in the states in $\mathcal{H}_{\text{kin}}$. This can be done by employing operators that correspond to easy to interpret geometric quantities. Prime examples for this are spatial areas and volumes. Indeed, it is possible to quantize areas and volumes with respect to the geometry on Σ on the Hilbert space $\mathcal{H}_{\text{kin}}$ [37.5–8, 51]. Since the quantum Einstein equations, in the form of the constraints, have not yet been taken into account, the physical implications of the results have to be considered with substantial care [37.52, 53]. There are, however, situations in which such quantities are observables, in the sense that they commute with the constraints. This is for example the case with the area of a black-hole horizon as considered in Sect. 37.4. In such cases the results that we are going to present have clear physical significance.

We consider first the case of areas: let S be a surface in Σ. When the field E is pulled back to S, one obtains a vector-valued two-form. The norm of this two-form is directly related to the area [37.54]

$$A_S = \int_S |E| \,. \tag{37.54}$$

This formula can be used as a starting point for quantization. Regularizing in terms of fluxes in the form of (37.28), substituting operators, and taking the regulator away leads to a well-defined, simple operator $\widehat{A}_S$. Its action on states with just a single edge is especially simple: if edge and surface do not intersect, the state is annihilated. If they do intersect once, one obtains

$$\widehat{A}_S \operatorname{Tr}[\pi_j(h_\alpha[A])] = 4\pi\beta\ell_{\mathrm{P}}^2\sqrt{j(j+1)}\operatorname{Tr}[\pi_j(h_\alpha[A])] \,. \tag{37.55}$$

Thus, these states are eigenstates of area, with the eigenvalue given as the square root of the eigenvalue of the SU(2)-Casimir operator in the representation given on the edge. A slightly more complicated action is obtained in the case of several intersections, and in particular if a vertex of the generalized spin network lies within the surface. Eigenstates and eigenvectors are nevertheless known explicitly. The full spectrum is of the form

$$a = 4\pi\beta\ell_{\mathrm{P}}^2 \sum_I \sqrt{\lambda_I} \,, \tag{37.56}$$

where the λ_I are half integers of the form

$$\lambda = 2j^{(u)}(j^{(u)}+1) + 2j^{(d)}(j^{(d)}+1) - j^{(u+d)}(j^{(u+d)}+1)\,, \tag{37.57}$$

with $j^{(u+d)}$ in $\{|j^{(u)} - j^{(d)}|, |j^{(u)} - j^{(d)}| + 1, \ldots, |j^{(u)} + j^{(d)}|\}$. As is seen in (37.56), the scale is set by the Planck area l_P^2. The eigenvalue density increases exponentially with area.

A similar procedure leads to an operator for volumes of subregions in Σ. This operator is substantially more complicated. Unlike the area operator, the action of which is purely in terms of the representation label of the edges, the volume operator acts on the vertices, by changing the maps that label them (*recoupling*). In fact, there are two slightly different versions of the volume operator [37.5, 6, 8], differing in the way the tangent-space structure of a vertex is taken into account. In either case, the spectrum is discrete, but not explicitly known. As in the case of the area operator, the volume operator is the result of a regularization procedure, and we will only state the result. Given a vertex v, one defines the operators

$$\widehat{q}_{e,e',e''} = \operatorname{sign}(\det(\dot{e}, \dot{e}', \dot{e}''))\epsilon^{ijk}\widehat{J}_i^{(e,v)}\widehat{J}_j^{(e',v)}\widehat{J}_k^{(e'',v)}\,,$$
$$\widehat{q}_v = \sum\nolimits_{e,e',e''\text{ at }v}\widehat{q}_{e,e',e''}\,. \tag{37.58}$$

The volume operator is then given by

$$\widehat{V} = \left(\left|\sum_v \widehat{q}_v\right|\right)^{1/2} \quad\text{or}\quad \widehat{V} = \left(\sum_v |\widehat{q}_v|\right)^{1/2}\,, \tag{37.59}$$

depending on which version, Ashtekar–Lewandowski (first expression) or Rovelli–Smolin (second one), one considers. A particular complication of the first version are the sign factors in the definition of $\widehat{q}_{e,e',e''}$, since these can have substantial influence on the spectrum for vertices of valence higher than three. While it may seem at first that arbitrary sign combinations may occur when letting e, e', e'' range over all the triples of edges at a given vertex, *Brunnemann* and *Rideout* [37.55] have observed that by no means all sign combinations can actually be realized by configurations of tangent vectors in $\mathbb{R}^3$.

Some remarkable analytic developments regarding the volume operator are given in [37.56–58], and a beautiful computer analysis of the lowest part of the spectrum can be found in [37.55, 59].

Thus, a picture emerges in which the vertices of spin networks can be associated with volumes, and the spins on the edges with the areas of the surfaces surrounding the volumes. This picture can be made even more detailed. The idea is to envision the geometry associated with the spin network as given by a gluing of polyhedra, one for each vertex.

To describe the aspects of this picture in more detail, it is useful to decompose the space of gauge-invariant states

$$\mathcal{H}_\gamma^0 = \bigoplus_{j\text{ admissible}} \mathcal{H}_{\gamma,j,l=0} \tag{37.60}$$

associated with a graph γ in a different way

$$\mathcal{H}_\gamma^0 = \bigoplus_{\{j_e|e\in\gamma\}} \bigotimes_{v\in\gamma} \operatorname{Inv}(j_e, e\text{ at }v)\,, \tag{37.61}$$

where $\operatorname{Inv}(j_e, e\text{ at }v)$ is the space of invariant tensors in the tensor product of representations at the vertex v. Thus, it is natural to investigate the geometric interpretation of the intertwiner spaces associated with the vertices. It turns out that it carries an action of $U(n_v)$, where n_v is the number of edges incoming at v [37.60]. Moreover, this action preserves the subspace of intertwiners with a fixed sum $\sum_e j_e$ of spins. This can be interpreted as the *preservation of area*, for the rotations of a quantized sphere.

The picture is rounded off by the following observations [37.61]: according to a theorem by Minkowski, there is a 1–1 correspondence between sets $\{a_i\}_{i=1,\ldots,n}$, $\{N_i\}_{i=1,\ldots,n}$ of positive real numbers and unit vectors in $\mathbb{R}^3$ satisfying the closure relation

$$\sum_i N_i a_i = 0 \tag{37.62}$$

and the equivalence classes of convex polyhedra under rotation. The a_i correspond to the areas of the faces, the N_i to their normal vectors under this correspondence. The space of this data carries a natural symplectic structure due to Kapovich and Millson, and its quantization can be precisely related to the structures at an n-valent vertex in LQG. To each edge e, one has a dual face whose area is given by the operator $|\boldsymbol{J}_e|$ and whose product of area and normal vector corresponds to $\boldsymbol{J}_e$. The closure condition (37.62) is then just gauge invariance at the vertex, and the intertwiners acquire an interpretation as the quantum states of a convex polyhedron. This has very interesting consequences for the quantum black hole [37.62], and one can show that the volume operator of loop quantum gravity is a quantization of the volume of the polyhedra [37.63].

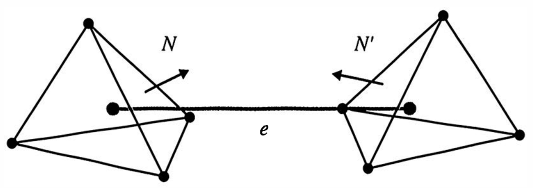


Fig. 37.6 The gluing of two flat tetrahedra dual to two vertices connected by an edge

One can finally consider how to combine the quantized polyhedra. The resulting geometries are called twisted geometries [37.64, 65]. The phase space associated with an edge e can be parameterized in a way that uses the geometry of two flat polyhedra glued together, possibly with extrinsic curvature. A point is given by (N, N', j, ξ), where N, N' are the unit normal vectors of the surfaces involved, and j describes their area; see Fig. 37.6. The operators $\widehat{J}^{(e,v)}$ can be understood as quantizations of jN, the parameter ξ is carrying the remaining information about the parallel transport from one polyhedron to the next, and is related to the holonomy operator in the quantum theory. An interesting point, corresponding to an earlier result [37.66], is that the geometries obtained in this way are twisted in the sense that the shapes of the glued triangles do not necessarily match.

37.3 Quantum Einstein Equations

In this section, we will examine how the constraints are formulated and implemented in loop quantum gravity. As we have indicated before, going over to the reduced phase space, by solving the constraints and going over to gauge orbits is classically equivalent to solving Einstein's equations. The information about the solutions is then contained in observables, i.e., functions that Poisson commute with all the constraints. In practice, to translate between a description in terms of observables and a more standard space–time description is very hard. Observables are by definition invariant under space–time diffeomorphisms, i.e., they must encode the information about space–time geometry in a very nonlocal way. For some discussion of how this can be done, see [37.67, 68].

In the quantum theory, to go from the unreduced to the physical theory, the only step required is the implementation of the constraints. In principle, this is done by constructing operators that correspond to the classical constraints, and restricting consideration to their joint kernel. The physical states are thus

$$\mathcal{H}_{\text{phys}} = \cap_\lambda \ker \widehat{C}_\lambda \,, \qquad (37.63)$$

where λ labels all the constraints, evaluated at all points. Observables then correspond to operators that commute with all the constraints, and can thus be restricted to $\mathcal{H}_{\text{phys}}$. This is the way that Einstein equations can be solved in the quantum theory.

37.3.1 Gauss Constraint

The simplest of the constraints (37.15) to implement is the Gauss constraint $G(A)$. Classically, it generates SU(2) gauge transformations which act on holonomies according to (37.31). There are two ways to do this in the quantum theory: one can either regularize the expression for G_I in terms of holonomies and fluxes, quantize the regularized expression, and remove the regulator to obtain a well-defined constraint operator in the limit. One can then determine the kernel of the quantum constraint.

Or, one can declare that all states in $\mathcal{H}_{\text{kin}}$ that are invariant under gauge transformations (37.31) are solutions to the constraint. Both strategies are viable, and lead to exactly the same result: the solution space $\mathcal{H}_{\text{gauge}}$ is a proper subspace of $\mathcal{H}_{\text{kin}}$, given by the subspaces with $l = 0$ in the decomposition (37.50) and (37.51). In other words, they are precisely the states labeled by spin networks [37.69, 70].

37.3.2 Diffeomorphism Constraint

The diffeomorphism constraint $C_a = E^b_I F^I_{ab}$ can not be quantized directly. One reason is that curvature can not be quantized on $\mathcal{H}_{\text{kin}}$, but one can see even on more general grounds that a quantization of C_a must run into difficulties: classically, this constraint generates the diffeomorphisms of Σ, and one expects the same of its quantum counterpart. Otherwise, one would have produced an anomalous implementation of the constraint, with possibly disastrous consequences for the theory. But, the diffeomorphisms ϕ of Σ already act on $\mathcal{H}_{\text{kin}}$, through unitary operators U_ϕ. These operators are, however, not strongly continuous in the diffeomorphisms (see (37.52)); in other words, they have no self-adjoint generators. Thus, C_a can not be directly

quantized without generating anomalies. But this is not a problem, because one understands what the gauge transformations generated by C_a are, and because they are acting in a simple manner on $\mathcal{H}_{\text{kin}}$. All one has to do is find states that are invariant under the action of the diffeomorphisms U_ϕ.

The action of the diffeomorphisms on cylindrical functions consists in moving the underlying graph

$$U_\varphi \Psi_\gamma = \Psi_{\varphi(\gamma)} \,. \tag{37.64}$$

Therefore, the only invariant state in $\mathcal{H}_{\text{gauge}}$ is the empty spin network. Rather than in $\mathcal{H}_{\text{diff}}$, the rest of the invariant states are lying in the dual of $\mathcal{H}_{\text{diff}}$. They can be found by group averaging. This procedure assigns to a state $\psi \in \mathcal{H}_{\text{gauge}}$ a diffeomorphism-invariant functional $\Gamma\psi$. The idea is to obtain invariant states, by averaging over the gauge group

$$(\Gamma\Psi)(\Psi') = (\text{Vol(Diff)})^{-1} \int_{\text{Diff}} D\varphi \langle U_\varphi \Psi | \Psi' \rangle \,. \tag{37.65}$$

To make this precise, the integration over the diffeomorphism group, and the division by its volume, have to be made sense of. These tasks would be hopeless, were it not for the unusual properties of the scalar product on $\mathcal{H}_{\text{kin}}$. In fact, the correct notion in this context of the integral over diffeomorphisms is that of a sum. A careful examination leads to the formula [37.22, 47]

$$\begin{aligned}(\Gamma\Psi_\gamma)(\Psi') = &\sum_{\varphi_1 \in \text{Diff}/\text{Diff}_\gamma} \frac{1}{|\text{GS}_\gamma|} \\ &\times \sum_{\varphi_2 \in \text{GS}_\gamma} \langle \varphi_1 * \varphi_2 * \Psi_\gamma | \Psi' \rangle \,.\end{aligned} \tag{37.66}$$

Here, Diff_γ is the subgroup of diffeomorphisms mapping γ onto itself, and TDiff_γ the subgroup of Diff which is the identity on γ. The quotient $\text{GS}_\gamma := \text{Diff}_\gamma/\text{TDiff}_\gamma$ is called the set of *graph symmetries*. It can be checked that this definition really gives diffeomorphism-invariant functionals over $\mathcal{H}_{\text{gauge}}$. An inner product can also be defined on these functionals, by setting

$$\langle \Gamma\Psi | \Gamma\Psi' \rangle = (\Gamma\Psi)(\Psi') \,. \tag{37.67}$$

Thus, one obtains a Hilbert space $\mathcal{H}_{\text{diff}}$ of gauge- and diffeomorphism-invariant quantum states.

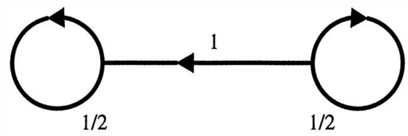


Fig. 37.7 The hourglass spin network gets mapped to zero under group averaging with respect to the diffeomorphism group

It is sometimes stated that diffeomorphism-invariant spin-network states are labeled by equivalence classes of spin networks under diffeomorphisms. This is a nice intuitive picture, but it is not entirely correct: the effects of (37.66) can be quite subtle. For example, the map Γ has a large kernel. Some spin networks, such as the *hourglass* (Fig. 37.7), are mapped to zero [37.71]. Diffeomorphism-invariant quantities can give rise to well-defined operators on $\mathcal{H}_{\text{diff}}$. An example is the total volume V_Σ of Σ. The corresponding operator on $\mathcal{H}_{\text{kin}}$ extends to $\mathcal{H}_{\text{diff}}$; thus, one obtains a well-defined notion of quantum volume. Areas of surfaces and volumes of subregions of Σ can similarly be quantized, provided surfaces and regions can be defined in a diffeomorphism-invariant fashion, for example by using a matter field as reference system.

37.3.3 Hamilton Constraint

As we have seen in Sect. 37.1, the Hamilton constraint of the classical theory is given by

$$\begin{aligned}C = &\underbrace{\frac{\beta}{2} \frac{1}{\sqrt{q}} E^a_i E^b_j \epsilon^{ij}{}_k F^k_{ab}}_{=:C_{\text{E}}} \\ &- \beta(1+\beta^2) \underbrace{\frac{1}{\sqrt{q}} E^a_i E^b_j K^i_{[a} K^j_{b]}}_{=:T} \,.\end{aligned} \tag{37.68}$$

In the present section, we will show how to turn this classical expression into a well-defined operator. The general difficulty with this is obviously that C is a complicated nonlinear function in the phase-space variables and hence ordering problems present themselves. There are also some specific difficulties with the expression:

- Equation (37.68) contains the inverse volume element. The volume element itself has a large kernel when quantized, so its inverse is ill defined.
- The expression (37.68) contains the curvature F of A, as well as the extrinsic curvature K. For neither

of them is there a simple operator in the quantum theory.

A guiding principle in the quantization process can be the Dirac algebra (37.18)–(37.20). In particular, the quantum Hamiltonian constraint should be invariant under gauge transformations, covariant under diffeomorphisms, and the commutator of two Hamilton constraints should give a diffeomorphism constraint.

We should say that the knowledge about the quantization and implementation of the Hamilton constraint is not complete. But, we will show that at least there is a strategy that leads to well-defined constraint operators. Given the difficulties outlined above, this is highly nontrivial in and of itself.

The quantization strategy we will describe in the following is due to *Thiemann* [37.72–74], but draws on important earlier work and ideas by Rovelli, Smolin, Lewandowski, and others. Our presentation follows that in [37.22].

Thiemann's Tricks

The quantization is based on two key ideas. The first one is to use various ingenious classical identities to express parts of the Hamilton constraint in terms of Poisson brackets before quantization. The second one is to express curvature in terms of holonomies.

Let

$$V = \int_\Sigma d^3x\sqrt{\det q}\,, \quad \overline{K} = \int_\Sigma d^3x K^i_a E^a_i \tag{37.69}$$

be the total volume of the spatial slice and the integrated extrinsic curvature. Then

$$\frac{E^a_i E^b_j \epsilon^{ijk}}{\sqrt{\det q}} = \frac{4}{\kappa}\epsilon^{abc}\{V, A^k_c\}\,, \quad K^j_a = \frac{2}{\kappa}\{\overline{K}, A^j_a\}\,. \tag{37.70}$$

These identities can be used to write

$$C_E(N) = c\int_\Sigma d^3x N\epsilon^{abc}\,\mathrm{Tr}\,(F_{ab}\{A_c, V\})\,, \tag{37.71}$$

$$T(N) = c'\int_\Sigma d^3x N\epsilon^{abc}\,\mathrm{Tr}\left(\{A_a,\overline{K}\}\{A_b,\overline{K}\}\{A_c,\overline{K}\}\right)\,, \tag{37.72}$$

where we have used the notation for the two parts of the Hamilton constraint introduced in (37.68). c and c' are simply two constants whose exact values are not so important for us. The idea behind these reformulations is that it is natural to replace Poisson brackets by commutators in the quantization process

$$\{\cdot,\cdot\} \longrightarrow \frac{1}{i\hbar}[\cdot,\cdot]\,. \tag{37.73}$$

This means that the quantization would be greatly simplified if operators existed for the quantities $V, \overline{K}$. Indeed, we have already seen in Sect. 37.2.4 that an operator exists for V. With respect to $\overline{K}$, the identity

$$\overline{K} = \{V, C_E\} \tag{37.74}$$

suggests to first quantize C_E, and then to use the commutator with the volume operator to define the operator for $\overline{K}$. Thus, we have already dealt with two of the difficulties regarding the quantization of C: the inverse volume element is gone, and the extrinsic curvature is dealt with. What remains is the quantization of the curvature F of A. Here, we use the well-known fact that holonomies encode information about curvature. Let S be an oriented surface such that the integral $\int_S F$ is small, and let α be the (oriented) boundary of S. Then the first term on the right- hand side of

$$\int_S F = \tfrac{1}{2}\left(h_\alpha - h_\alpha^{-1}\right) + O\left[\left(\int_S F\right)^2\right] \tag{37.75}$$

is a good approximation to the left-hand side. Let e be an edge starting at a point $s(e)$. A similar approximation plus a second Taylor expansion gives

$$\epsilon\dot{e}^a(s(e))\{A(s(e))_a, V\} \approx h_e^{-1}\{h_e, V\}\,, \tag{37.76}$$

where $\dot{e}$ is the tangent to e in a chosen parameterization $e(t)$, and ϵ is the coordinate length $\epsilon = \int_e dt$ of the edge in the given parameterization. In this way, we can express curvatures and connections by holonomies. Putting everything together, one obtains a Riemann-sum approximation of the Euclidean part of the constraint

$$\begin{aligned} C_E(N) &\approx C_E^{(\square)} \\ &:= c\sum_\square N(v_\square) \\ &\times \sum_{I=1}^{3}\left[\left(h^{-1}_{\alpha_I(\square)} - h_{\alpha_I(\square)}\right)h^{-1}_{s_I(\square)}\{h_{s_I(\square)}, V\}\right]\,. \end{aligned} \tag{37.77}$$

Here, $\{\Box\}$ is a decomposition of Σ into three-dimensional cells and, for each cell, a point $v_\Box$ has been fixed. $\{\alpha_I(\Box)\}$ is a set of loops and $\{s_I(\Box)\}$ a set of edges such that their tangents span the tangent space in the point $v_\Box$ in the following sense: there is a basis $\{b_I(\Box)\}$ of the tangent space at $v(\Box)$, such that $b_I(\Box)$ is tangent to both $\alpha_I(\Box)$ and $s_I(\Box)$, and compatible with their orientations. We call the data $(\{\Box\}, \{v_\Box\}, \{s_I(\Box)\}, \{\alpha_I(\Box)\})$ a *regulator* of C_E, and sometimes denote it simply by $\Box$. The exact shape of these cells, loops, and edges does not matter. The approximation is good as long as the cells are much smaller than the scale on which the fields A, E vary, and the loops and edges stay within the cells.

Finally, one can consider families of regulators such that the cells shrink to points. Then the corresponding approximations will converge to the exact result for a wide variety of such families.

The same kind of arguments can also be made for the second part of the Hamiltonian constraint $T(N)$. The connection components A_a in (37.72) can be replaced by holonomies along edges with suitable tangents, and the integrated exterior curvature $\overline{K}$ by Poisson brackets of V with the *regulated* Euclidean part (37.77), as per (37.74). The resulting expression is quite complicated and contains ambiguities, but the correct refinement limit is obtained for a large class of regulators.

Quantization

We will now come to the quantization. The general idea is clear: pick a family of regulators which converge to the continuum result. Replace Poisson brackets by commutators, and holonomies and volume operators by their operator counterparts, and obtain operators

$$\widehat{C}_E^{(\Box)}(N) = c\sum_{\Box} N(v_\Box) \times \sum_{I=1}^{3} \left(\left(h^{-1}_{\alpha_I(\Box)} - h_{\alpha_I(\Box)} \right) h^{-1}_{s_I(\Box)} \left[h_{s_I(\Box)}, \widehat{V} \right] \right) \tag{37.78}$$

on the kinematic Hilbert space. Now take the refinement limit $\Box \to \Sigma$ to obtain an operator $\widehat{C}_E$. There are, however, several difficulties when putting this program into practice:

1. In the limit of infinite refinement, the operator is in danger of creating infinitely many loops and edges. Hence, the limit may be ill defined.
2. Even if problem 1 can be overcome, the operator will generically not converge, since typically $\widehat{C}_E^{(\Box)}\Psi \perp \widehat{C}_E^{(\Box')}\Psi$ for regulators $\Box \neq \Box'$.
3. Since h and $\widehat{V}$ do not commute, there are ordering ambiguities.
4. There is a lot of ambiguity in the choice of regulators, since now there is no guarantee that different families of regulators will converge to the same operator, if they converge at all.

The first problem can be solved by a suitable ordering. Let us consider the action on a spin network. The volume operator acts only at the vertices; hence, ordering it to the right will force the loops and edges that are created by $\widehat{C}_E$ to be attached to the vertices of the spin network only. Thus, for a given spin network, only finitely many new edges and loops can be created. This also partially solves problem 3. To deal with the rest of the difficulties, we will be less ambitious, and not demand convergence in the kinematic Hilbert space. Rather, we consider the matrix elements of $\widehat{C}_E^{(\Box)}$ between one kinematic state and one diffeomorphism-invariant one. It turns out that due to the diffeomorphism invariance of the one state, many of the ambiguities in the attachment of the loops and edges do not change the matrix elements. What is more, for several types of regulators, it is known that the matrix elements converge

$$\lim_{\Box\to\Sigma} \left(\Psi | \widehat{C}_E^{(\Box)} | f_\gamma \right) \quad \text{is well defined} . \tag{37.79}$$

Typically, the matrix elements already become constant at a finite refinement, namely when the decomposition of Σ into cells is already so fine that there is at most one vertex of γ per cell.

Convergence of the above matrix elements does not imply that there exists a limit operator on the kinematic Hilbert space. Rather, we can interpret $(\Psi|\widehat{C}_E^{(\Box)}|$ as an element in the (algebraic) dual space of Cyl, and hence conclude that there is an operator

$$\widehat{C}_E^\dagger : \mathcal{H}_{\text{diff}} \longrightarrow \text{Cyl}^* . \tag{37.80}$$

The detailed features of this operator depend on the chosen family of regulators. But the generic features do not:

- $\widehat{C}_E^\dagger$ acts locally at the vertices.
- It acts by creating and annihilating edges and loops.

One can proceed in the same way with the quantization of $T(N)$, but, since the quantized expression contains double commutators with $\widehat{C}_E^{(\Box)}$, the operator action be-

comes extremely complicated. Nevertheless, it is well defined and finite.

Solutions

Given the definition of the Hamilton constraints we sketched above, what are the solutions? They are states Ψ in $\mathcal{H}_{\text{diff}}$ such that

$$\langle\Psi|C(N)f\rangle = 0 \quad \text{for all } f \in \text{Cyl and all } N\,. \tag{37.81}$$

One simple solution is the LQG vacuum $|\rangle$, which can also be interpreted as a state in $\mathcal{H}_{\text{diff}}$. But, more complicated solutions exist. For working out the set of solutions in some detail, details of the regularization used in the quantization of the constraints have to be fixed, since they do matter. Suffice it to say that so-called *exceptional edges* play an important role in the construction of solutions. Exceptional edges are edges of the type created by the quantum constraint itself. We will not discuss this in detail, but refer the reader to [37.22, 72–74] for more detailed accounts.

Solutions lie in the intersection of the kernels of all Hamilton constraints. Formally, the projector on this space can be expressed and approximated as follows [37.9]

$$\begin{aligned} P_C = \delta(\widehat{C}) &= \int \mathrm{D}N \mathrm{e}^{\mathrm{i}\widehat{C}(N)} \\ &= 1 + \mathrm{i}\int \mathrm{D}N \int N(X)\widehat{C}(x) \\ &\quad + \frac{\mathrm{i}^2}{2}\int \mathrm{D}N \iint N(X_1)N(X_2)\widehat{C}(x_1)\widehat{C}(x_2) + \dots\,. \end{aligned} \tag{37.82}$$

Here, $\widehat{C}(x)$ denotes the local action of the constraint, which is zero unless x is the position of a vertex of the state acted upon. The path integral over N gives an infinite result, but, by requiring diffeomorphism invariance, it can be split into a divergent term that can be normalized away, and a finite remainder [37.75].

The matrix elements of the projector can then be expanded into a series

$$\begin{aligned} \langle T_{\gamma_1}|P_C T_{\gamma_2}\rangle = \sum_{N=0}^{\infty}\sum_{v_1}\cdots\sum_{v_n} c_{v_1\dots v_N} \\ \times \left(T_{\gamma_1}|\widehat{C}(v_1)\widehat{C}(v_2)\dots\widehat{C}(v_N)|T_{\gamma_2}\right), \end{aligned} \tag{37.83}$$

where the finite sums are over all vertices of γ_2 and $c_{v_1\dots v_N}$ is the finite remainder of the integral over the lapse function. It only depends on the diffeomorphism equivalence class of the vertex set $\{v_1, v_2, \dots, v_N\}$. We note that a priori the multiple applications of the local constraint in (37.83) do not make sense, since we have up to now only defined the constraint operators in such a way that domain and range are disjoint; see (37.80). But, it is possible to enlarge the domain of definition in such a way that multiple applications of the constraints become possible [37.75, 76]. We will sketch how this is done when we discuss the question of anomalies below.

These matrix elements are interesting, because in principle they contain all the information about the inner product on the Hilbert space of physical states

$$\langle T_{\gamma_1}|P_C T_{\gamma_2}\rangle = \langle P_C T_{\gamma_1}|P_C T_{\gamma_2}\rangle_{\text{phys}}\,. \tag{37.84}$$

The expansion (37.83) can be interpreted as a kind of Feynman expansion, organized in terms of how many times the constraint acts. The individual terms can be nonzero only if the action of the constraint operators on T_{γ_2} produces exactly T_{γ_1}. Thus, the nonzero diagrams can be thought of as terms coming from the evolution of one spin network state into another. More precisely, they can be labeled by a two-complex, whose faces carry representations and whose edges carry intertwiners. The complex has the graphs γ_1, γ_2 as boundaries, and the internal vertices correspond to the action of the constraints. These diagrams are called *spin foams*, and they show up independently as spin-foam gravity, which is described in a separate chapter in this handbook. That they show up in an expression for the physical inner product of the canonical theory is a very encouraging link between canonical and covariant pictures. In fact, in the light of recent developments [37.10–16], one is getting close to actually having a precise correspondence

$$\begin{aligned} &\text{quantum Hamilton constraint} \\ &\longleftrightarrow \quad \text{spin-foam model}\,. \end{aligned} \tag{37.85}$$

We will now discuss some further aspects of the Hamilton constraint quantization.

Symmetry, Anomaly Freeness, and Ambiguities

In principle, it would be desirable to produce a symmetric, or even self-adjoint, Hamiltonian constraint

$$C^\dagger(N) = C(N)\,. \tag{37.86}$$

But, this turns out to be hard in practice, and there are even some no-go theorems [37.76]. Interestingly, there are heuristic arguments to the effect that one can not

have both symmetric constraints and a constraint algebra that is anomaly free.

We have seen that the constraints classically close to form an algebra with respect to the Poisson bracket. The same should happen on the quantum level, now with respect to the commutators. Otherwise, the gauge symmetries may have been broken when quantizing the theory. Such an anomaly in the gauge symmetries would strongly suggest the quantum theory to be unphysical. In particular, we are interested in the commutators

$$[C(M), C(N)] \,, \tag{37.87}$$

since, by the above construction, we can already see that the Hamilton constraints transform correctly under gauge transformations and diffeomorphisms. Classically the above commutator is proportional to a diffeomorphism constraint, hence at minimum one requires that the commutator should vanish states of $\mathcal{H}_{\text{diff}}$. The problem is that the constraints map $\mathcal{H}_{\text{diff}}$ to a certain subspace of Cyl* which is strictly larger than $\mathcal{H}_{\text{diff}}$. So, the above commutator is not well defined, as it stands. There are two proposed solutions to this problem. The first, by *Thiemann* [37.74], is to look at the commutator on $\mathcal{H}_{\text{kin}}$, before removing the regulator. He found that

$$\begin{aligned} &[C^{(\square)}(M), C^{(\square)}(N)] = \text{ something } \neq 0 \,, \\ &(\Psi| \text{ something } = 0 \text{ for } |\Psi\rangle \in \mathcal{H}_{\text{diff}} \,. \end{aligned} \tag{37.88}$$

In this sense

$$[C(M), C(N)]|_{\mathcal{H}_{\text{diff}}} = 0 \,, \tag{37.89}$$

and the quantization is anomaly free. The other solution to defining the commutator is by *Lewandowski* and *Marolf* [37.76]. They introduced a certain class of elements of Cyl* that is slightly larger than $\mathcal{H}_{\text{diff}}$. Without going into technical details, a *vertex-smooth state* $|\Psi\rangle$ is a state

$$|\Psi\rangle \in \text{Cyl}^* \ : \ (\Psi|U_\phi f_\gamma\rangle \text{ is a function of } V(\phi(\gamma)) \,, \tag{37.90}$$

i. e., of the set of vertices of the graph $\phi(\gamma)$, for *any* diffeomorphism ϕ. Trivial examples of vertex-smooth states are given by diffeomorphism-invariant states. A less trivial example is the linear functional given by

$$\Psi' \mapsto (\Psi| \int_\Sigma N \widehat{\sqrt{\det q}} || \Psi'\rangle \tag{37.91}$$

for a lapse function N and $|\Psi\rangle$ in $\mathcal{H}_{\text{diff}}$.

Lewandowski and *Marolf* observed that $(\Psi|C(N)$ is vertex smooth for a large class of regulators, and that its action can be extended to vertex-smooth states. Moreover, they found that

$$(\Psi_{\text{vs}}|[C(M), C(N)] = 0 \,, \tag{37.92}$$

where Ψ_{vs} is vertex smooth. As far as diffeomorphism-invariant states are concerned, this result would be expected for an anomaly-free representation. But, since it holds for all vertex-smooth states, it is surprising and a little worrisome, since the term in the Dirac algebra that results from the Poisson bracket of two Hamiltonian constraints, a diffeomorphism constraint, would be expected to act nontrivially on most vertex-smooth states. But, this has to be checked explicitly, and it may be possible to find quantizations of this term that indeed vanish on vertex-smooth states. New light on this question may be shed by new results of *Laddha* and *Varadarajan* [37.77–79], who employed new techniques to define constraints and their commutator algebra.

We should not finish without pointing out that there are various ambiguities in the above procedure that are poorly understood, for example regarding the loop attachment and the representation of the newly created links (see however [37.80]). Overall, it is however very encouraging that we can find a family of well-defined constraint operators that are anomaly free in a certain sense, and that lead to a convergence of the canonical and the spin-foam pictures. Given the complexity of the Hamiltonian constraints of general relativity, these results are highly nontrivial.

37.4 Black Holes

Black holes are fascinating objects predicted by general relativity. They even point beyond the classical theory, because of the singularities within, and because of the intriguing phenomenon of black-hole thermodynamics [37.81]. Therefore, they are a tempting subject of investigation in any theory of quantum gravity. Loop quantum gravity is able to successfully describe black-hole horizons in the quantum theory. Within this de-

scription, it is possible to identify degrees of freedom that carry the black-hole entropy, and prove, for a large class of black holes, the Bekenstein–Hawking area law.

The development of this subject is quite rich, with many turns and discussions as to the precise definition of the ensemble of quantum states; thus, our description will leave out many interesting aspects and references.

The first ideas were developed in [37.82]: black-hole entropy may be linked to topological quantum field theory, and [37.83]: spin-network edges pierce the horizon and endow it with area. The number of configurations of these edges (modulo diffeomorphisms) for a given total area is counted to obtain the entropy. A systematic and detailed treatment is in [37.84] (see also [37.85]), in which it was realized that the degrees of freedom on the horizon are described by a Chern–Simons theory and are essential in the calculation of the entropy. Reference [37.84] does contain errors in the state counting however, thereby wrongly concluding that only spin-network edges with spin 1/2 contribute significantly to the entropy counting. These errors were corrected in [37.86], where the horizon Hilbert space was correctly derived, its elements characterized in a combinatorial way, and the entropy calculation stated in combinatorial terms and partially carried out. It was also shown that the spin-1/2 edges are not generic, and a probability distribution for the edge spins derived. The combinatorial problem was fully solved in [37.87]. In [37.88, 89], it was assumed that a partial gauge fixing that had been used in [37.84] was unnecessary, and the ensuing combinatorial problem for the black-hole entropy was stated and solved. The area–entropy relation in the resulting more natural, but technically more challenging, setting was thus determined. In [37.90, 91], it was shown that dropping the partial gauge fixing as in [37.88, 89] can in fact be fully justified. This led to additional new insights [37.92]. In our description below, we will follow [37.90, 91].

There are interesting generalizations (for example [37.93, 94]) and modifications (for example [37.95–97]) of the formalism. Surprising fine structure has been found [37.98, 99] and analyzed [37.100–105]. The later works in this series are remarkable applications of number theory, statistics, and combinatorics.

The loop quantum gravity calculation does not start from solutions of the full theory. Rather, it quantizes gravity on a manifold with boundary Δ. In the simplest case, the boundary is assumed to be null, with topology $\mathbb{R}\times S^3$. Again, there are fields A and E on a manifold Σ, but now Σ has a boundary H. The boundary Δ is now required to be an *isolated horizon*, a quasi-local substitute for an event horizon. The phase-space description of a space–time with an isolated horizon can be worked out, and put in a form explicitly using connection variables. The details depend on the symmetry calls of the horizon. Results now exist for generic isolated horizons [37.94, 97]. Here, we will focus on spherically symmetric horizons for simplicity.

The isolated horizon imposes boundary conditions on the fields A and E at H

$$*E = -\frac{a_H}{\pi(1-\beta^2)}F(A)\,. \tag{37.93}$$

Here, a_H denotes the area of the horizon H. Furthermore, the symplectic structure acquires a surface term. The latter suggests, together with some technical aspects of the kinematical Hilbert space used in loop quantum gravity, quantizing the fields on the horizon separately from the bulk fields. The latter are quantized in the way described in Sect. 37.2. The only new aspect is that now edges of a spin network can end on the horizon. Such ends of spin-network edges are described by quantum numbers $m_p \in \{-j_p, -j_p+1, \dots, j_p-1, j_p\}$, where j_p is the representation label of the edge ending on the horizon, and p is a label for the end point (*puncture*). The quantum number represents the eigenvalue of the component of E normal to the horizon at the puncture.

The boundary term in the symplectic structure is that of a SU(2) Chern–Simons theory with level

$$k = \frac{a_H}{2\pi\beta(1-\beta^2)l_{\mathrm{P}}^2}\,, \tag{37.94}$$

and punctures where spin-network edges of the bulk theory end on the surface. The quantized Chern–Simons connection is flat, locally, but there are degrees of freedom at the punctures. These are – roughly speaking – described by quantum numbers s_p, m'_p, where the former is a half integer, and $m'_p \in \{-s_p, -s_p+1, \dots, s_p-1, s_p\}$. There is a constraint on the set of m'_p's coming from the fact that H is a sphere, and hence a loop going around all the punctures is contractible, and the corresponding holonomy must hence be trivial. The Hilbert space is equivalent to a subspace of the singlet component of the tensor product $\pi_{s_1}\otimes\pi_{s_2}\otimes\dots$ ranging over all punctures. The boundary condition (37.93) can be quantized to yield an operator equation. The solutions are tensor products of bulk and boundary states in which the quantum numbers (s_p, m'_p) and (j_p, m_p) are equal to each other at each puncture.

Now, if one fixes the quantum area of the black hole to be a, this bounds the number of punctures and the spins (j_p) labeling the representations. It becomes a rather complicated combinatorial problem to

determine the number $N(a)$ of quantum states with area a that satisfy the quantum boundary conditions. It was solved in [37.88, 89] and later, independently, in [37.106]. It turns out that

$$S(a) := \ln(N(a)) = \frac{\beta}{\iota_{\mathrm{SU}(2)}} \frac{a}{4\pi l_{\mathrm{P}}^2} - \frac{3}{2} \ln \frac{a}{l_{\mathrm{P}}^2} + O(a^0) \qquad (37.95)$$

as long as $\beta \leq \sqrt{3}$. Here, $\iota_{\mathrm{SU}(2)}$ is the constant that solves the equation

$$1 = \sum_{k=1}^{\infty} (k+1) \exp\left(-\frac{1}{2} \iota_{\mathrm{SU}(2)} \sqrt{k(k+2)}\right). \qquad (37.96)$$

One finds that $\iota_{\mathrm{SU}(2)} \approx 0.274$. One thus obtains the Bekenstein–Hawking area law upon setting $\beta = \iota_{\mathrm{SU}(2)}$.

37.5 Outlook

In this chapter, we have given an introduction to loop quantum gravity. In particular, we have discussed the Yang–Mills-type phase space that is its classical starting point; its quantization without the use of any kind of background structure; the quantum Riemannian geometry that results from it; the implementation of the constraints, i. e., the dynamics of general relativity; and the application of the theory to black holes. In this way, the chapter touched on at least some of the big achievements of loop quantum gravity, namely its description of quantum geometry and the corresponding dynamics, the quantum theory of (extrinsic and intrinsic) geometry, which comprises in particular geometric operators with a discrete spectrum, the scale of which is set by the Planck length, and diffeomorphism-invariant states. Based on this, well-defined Hamiltonian constraints can be obtained. This is a highly nontrivial result, given the complicated nature of the classical dynamics. Moreover, there is a clear connection to spin-foam gravity, which is invaluable since it opens the possibility of comparing results in an otherwise uncharted territory.

But, there are many omissions. For one thing, there was no discussion about loop quantum cosmology, the application of the techniques laid out in this chapter to symmetry-reduced sectors of general relativity. But this topic is covered in detail in another chapter of this handbook. Other omissions concern the quantization of matter fields along the same line as the geometry [37.45, 107–109]; and coherent states for quantum geometry [37.110–115].

A big omission are the many recent developments that could not be covered, among them:

- New quantization techniques have been developed that may allow for a check of the relations in the Dirac algebra in the quantum theory [37.77–79, 116]. This could be vital to assure an anomaly-free quantization.
- Spinorial variables and related techniques have been considered [37.115, 117–120], which may afford new insights for the dynamics.
- Progress has been made in identifying observables for general relativity [37.68, 121]. Matter systems have been identified that can be used as reference systems, to re-obtain a space–time picture in the canonical theory [37.122–125]
- First steps have been taken to treat quantum matter fields propagating on a quantized space–time, in an analogous approximation as used for quantum fields propagating on a classical background [37.20, 126, 127].
- A local notion of energy for isolated horizons [37.128, 129] allows us to investigate the thermodynamics of quantized horizons.
- Yang–Mills-type variables for higher dimensional gravity [37.40–43] and even supergravity [37.130] have been found. These allow the quantization techniques that were developed in loop quantum gravity to be applied to this much broader range of theories.

Let us finally list some important questions that are the subject of ongoing investigation in loop quantum gravity:

- Barbero–Immirzi parameter: what role does it ultimately play in loop quantum gravity with and without matter?
- Controlled approximations: loop quantum gravity is a nonperturbative approach to the quantization of gravity, but approximations will be vital to do physics. One question is how to find controlled approximations to situations with symmetries from the full theory. Another question is how to approximately solve the Hamilton constraints.
- Loop quantum gravity and matter: which types of matter can be consistently coupled to loop quan-

tum gravity? What are the implications of quantized space–time geometry to the propagation of matter?

- Physics from Hamilton constraints and Hamiltonians: how does one extract physics from the solutions to the constraints? In particular, one should be able to understand how ordinary quantum field theory and classical general relativity are embedded into loop quantum gravity. The first should correspond to a sector of quantum gravity where quantum fluctuations of the geometry are small but matter is still treated as a quantum object, whereas for general relativity both the matter and geometry quantum fluctuations are expected to be negligible. Furthermore, it is important to analyze how ambiguities in the quantization of constraints and physical Hamiltonians do reflect in physical properties of the theory.
- Connection to spin-foam gravity: what is the precise relation between scattering amplitudes and the physical inner product? Which quantization of the Hamilton constraint corresponds to which vertex amplitude?

For some of these, there are already insights. Answers to these questions will be crucial for the path that loop quantum gravity takes in the future.

References

37.1 A. Ashtekar: New variables for classical and quantum gravity, Phys. Rev. Lett. **57**, 2244 (1986)

37.2 A. Ashtekar: New variables for classical and quantum gravity, Phys. Rev. D **36**, 1587–1603 (1987)

37.3 R. Penrose: On the nature of quantum geometry (talk). In: *Magic Without Magic*, ed. by J.R. Klauder (W.H.Freemann, San Francisco 1972) pp. 333–354

37.4 C. Rovelli, L. Smolin: Loop space representation of quantum general relativity, Nucl. Phys. B **331**, 80 (1990)

37.5 C. Rovelli, L. Smolin: Discreteness of area and volume in quantum gravity, Nucl. Phys. B **442**, 593 (1995)

37.6 C. Rovelli, L. Smolin: Erratum, Nucl. Phys. B **456**, 753 (1995)

37.7 A. Ashtekar, J. Lewandowski: Quantum theory of geometry. 1: Area operators, Class. Quantum Gravity **14**, A55 (1997)

37.8 A. Ashtekar, J. Lewandowski: Quantum theory of geometry. 2. Volume operators, Adv. Theor. Math. Phys. **1**, 388 (1998)

37.9 M.P. Reisenberger, C. Rovelli: 'Sum over surfaces' form of loop quantum gravity, Phys. Rev. D **56**, 3490 (1997)

37.10 J. Engle, R. Pereira, C. Rovelli: The loop-quantum-gravity vertex-amplitude, Phys. Rev. Lett. **99**, 161301 (2007)

37.11 J. Engle, E. Livine, R. Pereira, C. Rovelli: LQG vertex with finite Immirzi parameter, Nucl. Phys. B **799**, 136 (2008)

37.12 W. Kaminski, M. Kisielowski, J. Lewandowski: Spin-foams for all loop quantum gravity, Class. Quantum Gravity **27**, 095006 (2010)

37.13 W. Kaminski, M. Kisielowski, J. Lewandowski: Erratum, Class. Quantum Gravity **29**, 049502 (2012)

37.14 W. Kaminski, M. Kisielowski, J. Lewandowski: The EPRL intertwiners, corrected partition function, Class. Quantum Gravity **27**, 165020 (2010)

37.15 W. Kaminski, M. Kisielowski, J. Lewandowski: Erratum, Class. Quantum Gravity **29**, 049501 (2012)

37.16 E. Alesci, T. Thiemann, A. Zipfel: Linking covariant and canonical LQG: New solutions to the Euclidean scalar constraint, Phys. Rev. D **86**, 024017 (2012)

37.17 M. Bojowald: Loop quantum cosmology, Living Rev. Relativ. **11**, 4 (2008)

37.18 M. Bojowald, G.M. Hossain, M. Kagan, S. Shankaranarayanan: Anomaly freedom in perturbative loop quantum gravity, Phys. Rev. D **78**, 063547 (2008)

37.19 A. Ashtekar, T. Pawlowski, P. Singh: Quantum nature of the big bang, Phys. Rev. Lett. **96**, 141301 (2006)

37.20 A. Ashtekar, W. Kaminski, J. Lewandowski: Quantum field theory on a cosmological, quantum space-time, Phys. Rev. D **79**, 064030 (2009)

37.21 W. Kaminski, J. Lewandowski, T. Pawlowski: Physical time and other conceptual issues of QG on the example of LQC, Class. Quantum Gravity **26**, 035012 (2009)

37.22 A. Ashtekar, J. Lewandowski: Background independent quantum gravity: A status report, Class. Quantum Gravity **21**, R53 (2004)

37.23 T. Thiemann: *Modern Canonical Quantum General Relativity* (Cambridge Univ. Press, Cambridge 2007)

37.24 C. Rovelli: Loop quantum gravity, Living Rev. Relativ. **11**, 5 (2008)

37.25 R. Gambini, J. Pullin: *A First Course in Loop Quantum Gravity* (Oxford Univ. Press, Oxford 2011)

37.26 R. Arnowitt, S. Deser, C.W. Misner: Canonical variables for general relativity, Phys. Rev. **117**(6), 1595–1602 (1960)

37.27 J.F.G. Barbero: Real Ashtekar variables for Lorentzian signature space times, Phys. Rev. D **51**, 5507 (1995)

37.28 G. Immirzi: Real and complex connections for canonical gravity, Class. Quantum Gravity **14**, L177 (1997)
37.29 S. Holst: Barbero's Hamiltonian derived from a generalized Hilbert–Palatini action, Phys. Rev. D **53**, 5966 (1996)
37.30 A. Perez, C. Rovelli: Physical effects of the Immirzi parameter, Phys. Rev. D **73**, 044013 (2006)
37.31 S. Alexandrov: SO(4,C) covariant Ashtekar–Barbero gravity and the Immirzi parameter, Class. Quantum Gravity **17**, 4255 (2000)
37.32 S. Alexandrov: On choice of connection in loop quantum gravity, Phys. Rev. D **65**, 024011 (2002)
37.33 S. Alexandrov, E.R. Livine: SU(2) loop quantum gravity seen from covariant theory, Phys. Rev. D **67**, 044009 (2003)
37.34 E.R. Livine: Projected spin networks for Lorentz connection: Linking spin foams and loop gravity, Class. Quantum Gravity **19**, 5525 (2002)
37.35 F. Cianfrani, G. Montani: Towards loop quantum gravity without the time gauge, Phys. Rev. Lett. **102**, 091301 (2009)
37.36 F. Cianfrani, G. Montani: The role of time gauge in quantizing gravity, Proceedings of the Third Stueckelberg Workshop on Relativistic Field Theories, ed. by N. Carlevaro, G.V. Vereshchagin, R. Ruffini (Cambridge Sc. Publishers, Cambridge 2010)
37.37 F. Cianfrani, G. Montani: The Immirzi parameter from an external scalar field, Phys. Rev. D **80**, 084040 (2009)
37.38 M. Geiller, M. Lachieze-Rey, K. Noui, F. Sardelli: A Lorentz-covariant connection for canonical gravity, SIGMA **7**, 083 (2011)
37.39 M. Geiller, M. Lachieze-Rey, K. Noui: A new look at Lorentz-covariant loop quantum gravity, Phys. Rev. D **84**, 044002 (2011)
37.40 N. Bodendorfer, T. Thiemann, A. Thurn: New variables for classical and quantum gravity in all dimensions I. Hamiltonian analysis, Class. Quantum Gravity **30**, 045001 (2013)
37.41 N. Bodendorfer, T. Thiemann, A. Thurn: New variables for classical and quantum gravity in all dimensions II. Lagrangian analysis, Class. Quantum Gravity **30**, 045002 (2013)
37.42 N. Bodendorfer, T. Thiemann, A. Thurn: New variables for classical and quantum gravity in all dimensions III. Quantum theory, Class. Quantum Gravity **30**, 045003 (2013)
37.43 N. Bodendorfer, T. Thiemann, A. Thurn: New variables for classical and quantum gravity in all dimensions IV. Matter coupling, Class. Quantum Gravity **30**, 045004 (2013)
37.44 S. Carlip: *Quantum Gravity in 2+1 Dimensions* (Cambridge Univ. Press, Cambridge 1998)
37.45 T. Thiemann: Kinematical Hilbert spaces for Fermionic and Higgs quantum field theories, Class. Quantum Gravity **15**, 1487 (1998)
37.46 A. Ashtekar, C.J. Isham: Inequivalent observable algebras: Another ambiguity in field quantization, Phys. Lett. B **274**, 393 (1992)
37.47 A. Ashtekar, J. Lewandowski, D. Marolf, J. Mourao, T. Thiemann: Quantization of diffeomorphism invariant theories of connections with local degrees of freedom, J. Math. Phys. **36**, 6456 (1995)
37.48 J. Lewandowski, A. Okolow, H. Sahlmann, T. Thiemann: Uniqueness of diffeomorphism invariant states on holonomy-flux algebras, Commun. Math. Phys. **267**, 703 (2006)
37.49 C. Fleischhack: Representations of the Weyl algebra in quantum geometry, Commun. Math. Phys. **285**, 67 (2009)
37.50 A. Ashtekar, J. Lewandowski: Projective techniques and functional integration for gauge theories, J. Math. Phys. **36**, 2170 (1995)
37.51 J. Lewandowski: Volume and quantizations, Class. Quantum Gravity **14**, 71 (1997)
37.52 B. Dittrich, T. Thiemann: Are the spectra of geometrical operators in loop quantum gravity really discrete?, J. Math. Phys. **50**, 012503 (2009)
37.53 C. Rovelli: Comment on: B. Dittrich, T. Thiemann: Are the spectra of geometrical operators in loop quantum gravity really discrete?, arXiv:0708.2481
37.54 C. Rovelli: Area is the length of Ashtekar's triad field, Phys. Rev. D **47**, 1703 (1993)
37.55 J. Brunnemann, D. Rideout: Properties of the volume operator in loop quantum gravity, I. Results, Class. Quantum Gravity **25**, 065001 (2008)
37.56 T. Thiemann: Closed formula for the matrix elements of the volume operator in canonical quantum gravity, J. Math. Phys. **39**, 3347 (1998)
37.57 J. Brunnemann, T. Thiemann: Simplification of the spectral analysis of the volume operator in loop quantum gravity, Class. Quantum Gravity **23**, 1289 (2006)
37.58 J. Brunnemann, D. Rideout: Oriented matroids – combinatorial structures underlying loop quantum gravity, Class. Quantum Gravity **27**, 205008 (2010)
37.59 J. Brunnemann, D. Rideout: Spectral analysis of the volume operator in loop quantum gravity, 11th Marcel Grossmann Meet. on Recent Developments in Theoretical and Experimental General Relativity, Gravitation and Relativistic Field Theories, Vol. MG11, ed. by H. Kleinert, R.T. Jantzen, R. Ruffini (World Scientific, Berlin 2006)
37.60 L. Freidel, E.R. Livine: The fine structure of SU(2) intertwiners from U(N) representations, J. Math. Phys. **51**, 082502 (2010)
37.61 E. Bianchi, P. Dona, S. Speziale: Polyhedra in loop quantum gravity, Phys. Rev. D **83**, 044035 (2011)
37.62 E. Bianchi: Black hole entropy, loop gravity, and polymer physics, Class. Quantum Gravity **28**, 114006 (2011)
37.63 E. Bianchi, H.M. Haggard: Discreteness of the volume of space from Bohr–Sommerfeld quantization, Phys. Rev. Lett. **107**, 011301 (2011)

37.64 L. Freidel, S. Speziale: Twisted geometries: A geometric parametrisation of SU(2) phase space, Phys. Rev. D **82**, 084040 (2010)
37.65 C. Rovelli, S. Speziale: On the geometry of loop quantum gravity on a graph, Phys. Rev. D **82**, 044018 (2010)
37.66 B. Dittrich, J.P. Ryan: Phase space descriptions for simplicial 4-D geometries, Class. Quantum Gravity **28**, 065006 (2011)
37.67 C. Rovelli: GPS observables in general relativity, Phys. Rev. D **65**, 044017 (2002)
37.68 B. Dittrich: Partial and complete observables for canonical general relativity, Class. Quantum Gravity **23**, 6155 (2006)
37.69 J.C. Baez: Spin network states in gauge theory, Adv. Math. **117**, 253 (1996)
37.70 C. Rovelli, L. Smolin: Spin networks and quantum gravity, Phys. Rev. D **52**, 5743 (1995)
37.71 C. Rovelli: *Quantum Gravity* (Cambridge Univ. Press, Cambridge 2004)
37.72 T. Thiemann: Quantum spin dynamics (QSD), Class. Quantum Gravity **15**, 839 (1998)
37.73 T. Thiemann: Quantum spin dynamics (QSD) 2, Class. Quantum Gravity **15**, 875 (1998)
37.74 T. Thiemann: QSD 3: Quantum constraint algebra and physical scalar product in quantum general relativity, Class. Quantum Gravity **15**, 1207 (1998)
37.75 C. Rovelli: Quantum gravity as a 'sum over surfaces', Nucl. Phys. Proc. Suppl. **57**, 28 (1997)
37.76 J. Lewandowski, D. Marolf: Loop constraints: A habitat and their algebra, Int. J. Mod. Phys. D **7**, 299 (1998)
37.77 A. Laddha, M. Varadarajan: Polymer quantization of the free scalar field and its classical limit, Class. Quantum Gravity **27**, 175010 (2010)
37.78 A. Laddha, M. Varadarajan: The Hamiltonian constraint in polymer parametrized field theory, Phys. Rev. D **83**, 025019 (2011)
37.79 A. Laddha, M. Varadarajan: The diffeomorphism constraint operator in loop quantum gravity, Class. Quantum Gravity **28**, 195010 (2011)
37.80 A. Perez: On the regularization ambiguities in loop quantum gravity, Phys. Rev. D **73**, 044007 (2006)
37.81 R.M. Wald: The thermodynamics of black holes, Living Rev. Relativ. **4**, 6 (2001)
37.82 L. Smolin: Linking topological quantum field theory and nonperturbative quantum gravity, J. Math. Phys. **36**, 6417 (1995)
37.83 K. Krasnov, C. Rovelli: Black holes in full quantum gravity, Class. Quantum Gravity **26**, 245009 (2009)
37.84 A. Ashtekar, J.C. Baez, K. Krasnov: Quantum geometry of isolated horizons and black hole entropy, Adv. Theor. Math. Phys. **4**, 1 (2000)
37.85 A. Ashtekar, J. Baez, A. Corichi, K. Krasnov: Quantum geometry and black hole entropy, Phys. Rev. Lett. **80**, 904 (1998)
37.86 M. Domagala, J. Lewandowski: Black hole entropy from quantum geometry, Class. Quantum Gravity **21**, 5233 (2004)
37.87 K.A. Meissner: Black hole entropy in loop quantum gravity, Class. Quantum Gravity **21**, 5245 (2004)
37.88 R.K. Kaul, P. Majumdar: Quantum black hole entropy, Phys. Lett. B **439**, 267 (1998)
37.89 R.K. Kaul, P. Majumdar: Logarithmic correction to the Bekenstein–Hawking entropy, Phys. Rev. Lett. **84**, 5255 (2000)
37.90 J. Engle, A. Perez, K. Noui: Black hole entropy and SU(2) Chern–Simons theory, Phys. Rev. Lett. **105**, 031302 (2010)
37.91 J. Engle, K. Noui, A. Perez, D. Pranzetti: Black hole entropy from an SU(2)-invariant formulation of type I isolated horizons, Phys. Rev. D **82**, 044050 (2010)
37.92 E. Bianchi: Black hole entropy, loop gravity, and polymer physics, Class. Quantum Gravity **28**, 114006 (2011)
37.93 A. Ashtekar, J. Engle, C. Van Den Broeck: Quantum horizons and black hole entropy: Inclusion of distortion and rotation, Class. Quantum Gravity **22**, L27 (2005)
37.94 C. Beetle, J. Engle: Generic isolated horizons in loop quantum gravity, Class. Quantum Gravity **27**, 235024 (2010)
37.95 A. Ghosh, P. Mitra: Counting black hole microscopic states in loop quantum gravity, Phys. Rev. D **74**, 064026 (2006)
37.96 G.J. Fernando Barbero, J. Lewandowski, E.J.S. Villasenor: Flux-area operator and black hole entropy, Phys. Rev. D **80**, 044016 (2009)
37.97 A. Perez, D. Pranzetti: Static isolated horizons: SU(2) invariant phase space, quantization, and black hole entropy, Entropy **13**, 744 (2011)
37.98 A. Corichi, J. Diaz-Polo, E. Fernandez-Borja: Black hole entropy quantization, Phys. Rev. Lett. **98**, 181301 (2007)
37.99 A. Corichi, J. Diaz-Polo, E. Fernandez-Borja: Quantum geometry and microscopic black hole entropy, Class. Quantum Gravity **24**, 243 (2007)
37.100 J. Diaz-Polo, E. Fernandez-Borja: Note on black hole radiation spectrum in Loop Quantum Gravity, Class. Quantum Gravity **25**, 105007 (2008)
37.101 H. Sahlmann: Entropy calculation for a toy black hole, Class. Quantum Gravity **25**, 055004 (2008)
37.102 H. Sahlmann: Toward explaining black hole entropy quantization in loop quantum gravity, Phys. Rev. D **76**, 104050 (2007)
37.103 I. Agullo, E.F. Borja, J. Diaz-Polo: Computing black hole entropy in loop quantum gravity from a conformal field theory perspective, J. Cosmol. Astropart. Phys. **0907**, 016 (2009)
37.104 I. Agullo, J. Fernando Barbero, E.F. Borja, J. Diaz-Polo, E.J.S. Villasenor: Detailed black hole state counting in loop quantum gravity, Phys. Rev. D **82**, 084029 (2010)

37.105 G.J. Fernando Barbero, E.J.S. Villasenor: Statistical description of the black hole degeneracy spectrum, Phys. Rev. D **83**, 104013 (2011)

37.106 I. Agullo, G.J. Fernando Barbero, E.F. Borja, J. Diaz-Polo, E.J.S. Villasenor: The combinatorics of the SU(2) black hole entropy in loop quantum gravity, Phys. Rev. D **80**, 084006 (2009)

37.107 J.C. Baez, K.V. Krasnov: Quantization of diffeomorphism invariant theories with fermions, J. Math. Phys. **39**, 1251 (1998)

37.108 T. Thiemann: QSD 5: Quantum gravity as the natural regulator of matter quantum field theories, Class. Quantum Gravity **15**, 1281 (1998)

37.109 A. Ashtekar, J. Lewandowski, H. Sahlmann: Polymer and Fock representations for a scalar field, Class. Quantum Gravity **20**, L11 (2003)

37.110 T. Thiemann: Gauge field theory coherent states (GCS): 1. General properties, Class. Quantum Gravity **18**, 2025 (2001)

37.111 T. Thiemann: Complexifier coherent states for quantum general relativity, Class. Quantum Gravity **23**, 2063 (2006)

37.112 B. Bahr, T. Thiemann: Gauge-invariant coherent states for loop quantum gravity. II. Non-Abelian gauge groups, Class. Quantum Gravity **26**, 045012 (2009)

37.113 C. Flori, T. Thiemann: Semiclassical analysis of the loop quantum gravity volume operator. I. Flux coherent states, arXiv:0812.1537 [gr-qc]

37.114 C. Flori: Semiclassical analysis of the loop quantum gravity volume operator: Area coherent states, arXiv:0904.1303 [gr-qc]

37.115 L. Freidel, E.R. Livine: U(N) coherent states for loop quantum gravity, J. Math. Phys. **52**, 052502 (2011)

37.116 C. Tomlin, M. Varadarajan: Towards an anomaly-free quantum dynamics for a weak coupling limit of Euclidean gravity, Class. Quantum Gravity **87**, 044039 (2013)

37.117 E.F. Borja, L. Freidel, I. Garay, E.R. Livine: U(N) tools for loop quantum gravity: The return of the spinor, Class. Quantum Gravity **28**, 055005 (2011)

37.118 E.R. Livine, J. Tambornino: Spinor representation for loop quantum gravity, J. Math. Phys. **53**, 012503 (2012)

37.119 E.R. Livine, S. Speziale, J. Tambornino: Twistor networks and covariant twisted geometries, Phys. Rev. D **85**, 064002 (2012)

37.120 E.R. Livine, J. Tambornino: Loop gravity in terms of spinors, J. Phys. Conf. Ser. **360**, 012023 (2012)

37.121 B. Dittrich, J. Tambornino: A Perturbative approach to Dirac observables and their space-time algebra, Class. Quantum Gravity **24**, 757 (2007)

37.122 K. Giesel, S. Hofmann, T. Thiemann, O. Winkler: Manifestly gauge-invariant general relativistic perturbation theory, I. Foundations, Class. Quantum Gravity **27**, 055005 (2010)

37.123 M. Domagala, K. Giesel, W. Kaminski, J. Lewandowski: Gravity quantized: Loop quantum gravity with a scalar field, Phys. Rev. D **82**, 104038 (2010)

37.124 K. Giesel, T. Thiemann: Algebraic quantum gravity (AQG). IV. Reduced phase space quantisation of loop quantum gravity, Class. Quantum Gravity **27**, 175009 (2010)

37.125 V. Husain, T. Pawlowski: Time and a physical Hamiltonian for quantum gravity, Phys. Rev. Lett. **108**, 141301 (2012)

37.126 H. Sahlmann, T. Thiemann: Towards the QFT on curved space-time limit of QGR. 1. A general scheme, Class. Quantum Gravity **23**, 867 (2006)

37.127 H. Sahlmann, T. Thiemann: Towards the QFT on curved space-time limit of QGR. 2. A concrete implementation, Class. Quantum Gravity **23**, 909 (2006)

37.128 A. Ghosh, A. Perez: Black hole entropy and isolated horizons thermodynamics, Phys. Rev. Lett. **107**, 241301 (2011)

37.129 A. Ghosh, A. Perez: Erratum, Phys. Rev. Lett. **108**, 169901 (2012)

37.130 N. Bodendorfer, T. Thiemann, A. Thurn: Towards loop quantum supergravity (LQSG), Phys. Lett. B **711**, 205 (2012)

38. Spin Foams

Jonathan S. Engle

The spin foam framework provides a way to define the dynamics of canonical loop quantum gravity in a spacetime covariant way, by using a path integral over histories of quantum states which can be interpreted as *quantum space-times*. This chapter provides a basic conceptual introduction to spin foams as well as a view of some current research topics.

Ever since special relativity, space and time have become seamlessly merged into a single entity, and space–time symmetries, such as Lorentz invariance, have played a key role in our fundamental understanding of nature. Quantum mechanics, however, did not originally conform to this new way of thinking. The original formulation of quantum mechanics, called *canonical*, involves wavefunctions, operators, Hamiltonians, and time evolution in a way that treats time very differently from space. This situation was improved by Feynman, who formulated quantum mechanics in terms of probabilities calculated by summing over amplitudes associated with classical histories – the path-integral formulation of quantum mechanics. As histories are naturally space–time objects in which space and time can be viewed *on equal footing*, the path-integral formulation allowed, for the first time, space–time symmetries to be manifest in a general quantum theory.

The key insight of Einstein's theory of gravity, general relativity, is that gravity is space–time geometry. Space–time geometry, the one *background structure* – i. e., nondynamical space–time structure – remaining

after special relativity, was discovered to be dynamical and to describe the gravitational field, revealing nature to be *background independent*. Background independence can equivalently be expressed in terms of a profound enlargement of the basic space–time symmetry group of physics: invariance under Lorentz transformations and translations is replaced by invariance under the much larger group of *space–time diffeomorphisms*.

We have already seen in the chapter by Sahlmann on gravity, geometry, and the quantum, a canonical quantization of Einstein's gravity, and hence of geometry, in which geometric operators are derived with discrete eigenvalues [38.1–3]. Instead of space being a smooth continuum, we see that it comes in discrete quanta – minimal *chunks of space*. Furthermore, as discussed in the chapter by Agullo and Corichi, when applied to cosmology, this quantum theory of gravity leads to a new understanding of the Big Bang in which usually problematic infinities are resolved, and one can actually ask what happened *before* the Big Bang. In spite of these successes, because it is a canonical theory, it has as a drawback that space–time symmetries, in particular space–time diffeomorphism symmetries, are not manifest. Equivalently, the preferred separation between space and time prevents full background independence from being manifest.

One can ask: is there a way to construct a path-integral formulation of quantum gravity, in which the most radical discovery of general relativity, background independence or, equivalently, space–time diffeomorphism invariance, can be fully manifest, which nevertheless retains the successes of the canonical theory? This is the question leading to the spin-foam program. In answering it one must understand more carefully the relationship between the canonical and path-integral formulations of quantum mechanics, and in particular how these apply to general relativity, with its special subtleties such as the *problem of time* discussed in Chap. 33 and Chap. 36. The end result is a path integral in which, instead of summing over classical space–time histories, one sums over *histories of quantum states of space*. These histories have a natural space–time interpretation and thus may be thought of as *quantum space–times*. The resulting sum over histories then provides a framework for defining the *dynamics* of loop quantum gravity (LQG) in which space and time are unified, in the spirit of special and general relativity. Due to their structure and the way they are labeled, these *quantum space–times* have been named *spin foams* by *Baez* [38.4], a name which thenceforth has been used to refer to the entire program.

In this chapter we hope to give the reader a broad view of the conceptual ideas behind spin foams, the ideas that have led to the spin-foam model currently most often used in the community, as well as provide a view of current avenues of investigation. For a more detailed, complete review of spin foams [38.5] is recommended to the interested reader.

38.1 Background Ideas

38.1.1 The Path Integral as a Sum over Histories of Quantum States

The first formulation of quantum mechanics that was discovered, and that one learns, is the canonical formulation. We review here briefly the basic structure of a canonical quantum theory. The possible states of a canonical quantum system form a *vector space*, that is, they are such that states can be rescaled by real numbers and added to each other. Additionally, one has an *inner product*, which assigns to every two states ϕ and ψ a complex number $\langle\psi,\phi\rangle$, which may be roughly thought of as the *overlap* between states ϕ and ψ. A vector space equipped with such an inner product is called a *Hilbert space*; one often uses the phrase *the Hilbert space of quantum states*. For each possible measurable quantity, such as position, momentum, angular momentum, or energy – or in the case of general relativity, areas of surfaces and volumes of regions – there is a corresponding operator $\hat{O}$ mapping states to states. A number λ is a possible outcome of a measurement of $\hat{O}$ only if there exists a state ψ such that $\hat{O}\psi = \lambda\psi$. When the state of the system is ψ, then a measurement of $\hat{O}$ yields λ with certainty. Such a λ and corresponding ψ are called an *eigenvalue* and an *eigenstate* of $\hat{O}$. The set of all possible eigenvalues – and hence possible results of a measurement – of $\hat{O}$ is called the *spectrum* of $\hat{O}$. Depending on the operator, its spectrum may include all real numbers, or it may only include a discrete set of possible numbers. This is the source of the name *quantum*: that some quantities, when measured, can only come in discrete increments, called quanta.

Time evolution in canonical quantum theory is determined by Schrödinger's equation

$$i\hbar \frac{d\psi}{dt} = \hat{H}\psi \,, \tag{38.1}$$

where $\hbar$ is Planck's constant divided by 2π, and $\hat{H}$ is the *Hamiltonian* operator, which corresponds to the total energy of the system. If the system starts in an initial state $\psi(t_i)$, Schrödinger's equation will uniquely determine its state $\psi(t_f)$ at any later time $t_f = t_i + T$, thus providing a map $U(T)$ from possible initial states $\psi(t_i)$ to final states $\psi(t_f)$, called a *time-evolution map*. Using the time-evolution map, and given two states ψ_i, ψ_f, and two times t_i, t_f, one can define a quantity

$$\mathcal{A}(\psi_f, t_f; \psi_i, t_i) := \langle \psi_f, U(t_f - t_i)\psi_i \rangle \,,$$

called a *transition amplitude*. The transition amplitude is of direct use for making predictions: if the system is prepared in an initial state ψ_i at time t_i, the transition amplitude tells us the probability of measuring the system to be in a final state ψ_f at time t_f. (Specifically, this probability is given by the formula $|\mathcal{A}(\psi_f, t_f; \psi_i, t_i)|^2 / |\langle \psi_f, \psi_f \rangle \langle \psi_i, \psi_i \rangle|$.)

The transition amplitude contains all information about the dynamics of the quantum system. At the heart of the path-integral formulation of quantum mechanics is Feynman's insight that the transition amplitude can be *rewritten* in terms of purely classical, *space–time* quantities. Consider, for example, a single free particle, and consider the case in which ψ_i and ψ_f are *eigenstates of position*, i. e., states in which the position of the particle is exactly defined, being equal to some x_i and x_f, respectively. We write $\psi_i = |x_i\rangle$ and $\psi_f = |x_f\rangle$. In this case, one usually uses a simpler notation for the transition amplitude: $\mathcal{A}(x_f, t_f; x_i, t_i) := \mathcal{A}(|x_f\rangle, t_i; |x_i\rangle, t_i)$. The expression for the transition amplitude can be rewritten as

$$\mathcal{A}(x_f, t_f; x_i, t_i) = \langle x_f, U(\tfrac{T}{N}) \cdots U(\tfrac{T}{N}) U(\tfrac{T}{N}) U(\tfrac{T}{N}) x_i \rangle \,, \tag{38.2}$$

where $T := t_f - t_i$ and one has used the fact that the time evolution $U(T)$ is equivalent to performing N evolutions over the smaller time T/N. The eigenstates of position $|x\rangle$ satisfy the following identity: for all $\psi, \phi \in \mathcal{H}$

$$\langle \psi, \phi \rangle = \int_{-\infty}^{\infty} \langle \psi, x \rangle \langle x, \phi \rangle \, dx \,. \tag{38.3}$$

This is known as a *completeness relation* or *resolution of the identity*. Note that the range of integration on the right-hand side includes all possible values which can result from a measurement of the position $\hat{x}$ – that is, the integral is over the spectrum of $\hat{x}$. If $\hat{x}$ were *quantized*, that is, if its spectrum were discrete, this integral would be replaced by a *sum* over the discrete spectrum. We will remark on this later. Applying the identity (38.3) to (38.2) $N - 1$ times, in sequence, one obtains

$$\begin{aligned}
\mathcal{A}(x_f, t_f; x_i, t_i) &= \int \langle x_f, U(\tfrac{T}{N}) \cdots U(\tfrac{T}{N}) U(\tfrac{T}{N}) x_1 \rangle \\
&\quad \times \langle x_1, U(\tfrac{T}{N}) x_i \rangle \, dx_1 \\
&= \iint \langle x_f, U(\tfrac{T}{N}) \cdots U(\tfrac{T}{N}) x_2 \rangle \\
&\quad \times \langle x_2, U(\tfrac{T}{N}) x_1 \rangle \langle x_1, U(\tfrac{T}{N}) x_i \rangle \, dx_1 \, dx_2 \\
&\;\;\vdots \\
&= \iint \cdots \int \langle x_f, U(\tfrac{T}{N}) x_{N-1} \rangle \\
&\quad \cdots \langle x_2, U(\tfrac{T}{N}) x_1 \rangle \langle x_1, U(\tfrac{T}{N}) x_i \rangle \\
&\quad \times dx_1 \, dx_2 \cdots dx_{N-1} \,.
\end{aligned}$$

In this expression one has introduced $N - 1$ intermediate position eigenstates, and one integrates over all possible such intermediate states. This sequence of intermediate states forms a discrete history of quantum states. Note that the above expression is exact for any N. If one takes the limit at N approaches infinity, the discrete histories are replaced by continuum histories of quantum states, and one obtains the *path integral*;

$$\mathcal{A}(x_f, t_f; x_i, t_i) = \int_{\substack{x(t_i)=x_i \\ x(t_f)=x_f}} \exp\left(\frac{i}{\hbar} S[x(\cdot)]\right) \mathcal{D}x(\cdot) \,, \tag{38.4}$$

where, heuristically, $\mathcal{D}x(\cdot)$ denotes $\prod_t dx(t)$, and $S[x(\cdot)]$ is the *action* for the theory. The action is a purely classical quantity, which specifies a number for each possible classical history $x(t)$. It is maximized or minimized when $x(t)$ is a *solution to the classical equations of motion*. Because of the close relation between integrals and sums (one is just a limit of the other), the integral in (38.4) is also loosely referred to as a *sum* over paths, or a *sum* over histories. If the position operator $\hat{x}$ had had a *discrete* spectrum, so that only a discrete set of values were allowed for x, as already mentioned, the resolution

of the identity (38.3) would have actually been replaced by a sum, and the final path integral (38.4) would have actually become a sum rather than an integral. There are also cases where the final expression for the propagator (38.4) involves a combination of sums and integrals. In this chapter, as in much of the literature on path integrals, we will be loose with the distinction between sums over paths and integrals over paths, and will use the terms *path integral*, *sum over paths*, and *sum over histories* interchangeably. Nevertheless, because, in the case of quantum gravity, the primary interest of this chapter, one will turn out to have mostly sums, we will generally prefer to use the term sum.

Equation (38.4) provides an expression for the transition amplitude from a position x_i at time t_i to a position x_f at time t_f, an expression that involves only a sum over *classical paths* $x(t)$ that start at x_i and end at x_f, and the *classical* action $S[x(t)]$ depending on this path. The canonical theory enters into the expression in *one way only*: it determines the spectrum of $\hat{x}$ and hence the allowed values that the history of eigenvalues $x(t)$ can take at each moment in time. Other than this, *classical* physics is the only input for this expression. Because of this, Feynman made the radical proposal that this formula, which encodes all physical predictions for the system in question, be a new starting point for the very definition of the quantum theory.

However, care is necessary. As noted, one piece of information from the canonical quantum theory *does* remain: it is the *canonical* theory which tells us the spectrum of the position operator $\hat{x}$, and hence the possible positions which one sums over in the path integral. In the case of the free particle, the spectrum of the position includes all real numbers, so that, in fact, the sum is equivalent to a sum over all classical histories. However, in other theories this is not necessarily the case. In particular, in the case of gravity, one must sum over *histories of geometry*. But, one of the seminal results of loop quantum gravity is that *geometry is quantized*. Areas of surfaces and volumes of regions can only take on discrete sets of possible values. Thus, one should not sum over all histories of *classical* geometries, but rather over histories of the allowable *quantum geometries* predicted by loop quantum gravity. This is the insight leading to the spin-foam program.

Before closing this section, let us remark that the integrand in (38.4) can be interpreted as giving the probability amplitude for a *single history* $x(t)$

$$\mathcal{A}[x(\cdot)] = \exp\left(\frac{i}{\hbar}S[x(\cdot)]\right) . \qquad (38.5)$$

The total transition amplitude (38.4) is then obtained by integrating (or adding) the amplitudes (38.5) associated with all histories compatible with the relevant *boundary conditions*, $x(t_i) = x_i$, $x(t_f) = x_f$.

The precise form (38.5) for the amplitude of each history not only arises from the canonical quantum theory in the manner presented above, but it is also important for the correct *classical limit* of the quantum theory. When constructing a quantum theory, usually the corresponding classical theory is already well tested experimentally. In order to be consistent with known experiments, it is therefore crucial that the predictions of the quantum theory agree with those of the classical theory in situations where the effects of quantum mechanics can be neglected. One way of stating this requirement is that if appropriate combinations of the physical scales in the situation are large compared to Planck's constant (so that Planck's constant can effectively be scaled to zero), then the quantum theory should yield the same predictions as the corresponding classical theory. The limit here described – that of either large physical scales or Planck's constant being scaled to zero – is what is called the *classical limit* of a quantum theory, and the requirement that this yield predictions equivalent to the classical theory is called the requirement of having the *correct classical limit*.

Let us consider the classical limit of the path integral. For the present argument, it is easiest to cast this as the limit in which *Planck's constant is scaled to zero*. In this limit, the phase $\frac{1}{\hbar}S$ of the amplitude (38.5) becomes very large compared to 2π. If one divides up the domain of integration – the space of histories compatible with the boundary conditions – into many small neighborhoods, one finds that, in the vast majority of these neighborhoods, the phase of the integrand will oscillate very fast. As a consequence, in such neighborhoods, there tend to be an equal number of opposite phase contributions from the path integral which cancel each other, so that the total contribution from such neighborhoods tends to be zero (Fig. 38.1). The only neighborhoods where the phase is *not* oscillating fast are those where $S[x(\cdot)]$ does not change very much when $x(\cdot)$ changes. These are precisely the neighborhoods where $S[x(\cdot)]$ is *maximum* or *minimum*, that is, precisely the neighborhoods containing a *solution to the classical equations of motion*. Thus, one sees that, in the classical limit, only histories near solutions to the classical equations of motion contribute to the path integral. This is key to obtaining the correct classical limit of the quantum theory. The Feynman prescription (38.5) for the

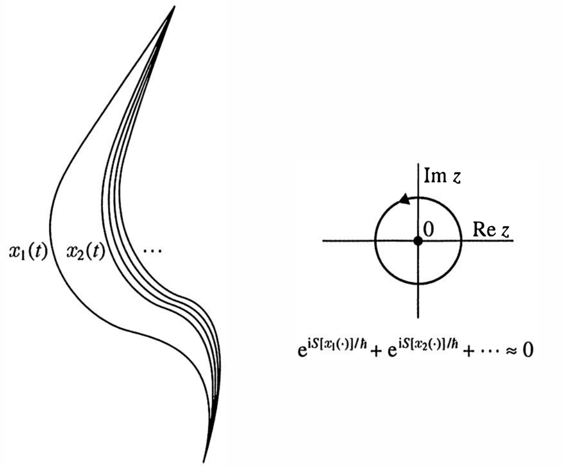


Fig. 38.1 In the classical limit, at histories $x(\cdot)$ where $(1/\hbar)S[x(\cdot)]$ changes, it changes very fast, so that the phases from the sum over histories near $x(\cdot)$ tend to cancel. When $S[x(\cdot)]$ does *not* change with $x(\cdot)$, the phases do not cancel, but reinforce each other. This happens when $x(\cdot)$ is a local minimum or maximum of $S[x(\cdot)]$ – that is, when $x(\cdot)$ is a solution to the classical equations of motion. In this way the classical solutions dominate the sum over histories in the classical limit

probability amplitude of a single history is thus directly related to ensuring the correct classical limit.

38.1.2 Field Theory and the General Boundary Formulation of Quantum Mechanics

Before going on to the specific case of gravity, we take the opportunity to first discuss field theory, and introduce what is known as the *general boundary* formulation of quantum mechanics [38.2, 6, 7]. In the case of field theory, instead of integrating over possible paths $x(t)$ of a particle from time $t = t_i$ to $t = t_f$ as in (38.4), one integrates over possible *fields* $\phi(x)$ on the *four-dimensional space–time region* bounded by the instants $t = t_i$ and $t = t_f$. In the general boundary formulation, this region is allowed to be replaced by *any* space–time region. The biggest advantage of this formulation of quantum mechanics is that, by choosing this region to be finite, it permits purely local calculations in a quantum field theory in which one need not worry about the asymptotic behavior of states at infinity. Not only are such calculations more consistent with the locality of the measuring apparatus one would actually use, but

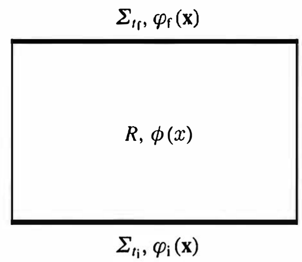


Fig. 38.2 In the path integral for the scalar field, one sums over all *fields* $\phi(x)$ on some *space–time region R* compatible with given initial values $\varphi_i(\mathbf{x})$ on the initial hypersurface Σ_{t_i} and final values $\varphi_f(\mathbf{x})$ on the final hypersurface Σ_{t_f}, where Σ_{t_i} and Σ_{t_f} bound R

they are technically simpler, and have been central to most work in spin foams up until now.

The Free Scalar Field

As an example, let us look at the case of a scalar field in Minkowksi space. In this case, one has as basic canonical variables $\varphi(\mathbf{x})$ and its conjugate momentum field $\pi(\mathbf{x})$, and corresponding operators $\hat{\varphi}(\mathbf{x})$, $\hat{\pi}(\mathbf{x})$. We here use bold to denote spatial points. One has a complete set of simultaneous eigenstates $|\varphi(\mathbf{x})\rangle$ of the operators $\hat{\varphi}(\mathbf{x})$, each now labeled by a *field* $\varphi(\mathbf{x})$ on space. A history of such fields, $\phi(t, \mathbf{x}) = \phi(x)$, is a field on the four-dimensional *space–time region R* bounded by the three-dimensional *instant-time* hypersurfaces $t = t_i$ and $t = t_f$, which shall be denoted Σ_{t_i} and Σ_{t_f}, respectively (Fig. 38.2). Note that space–*time* points such as x will not be bolded. Equation (38.4) becomes, in this case

$$\mathcal{A}^{\text{scalar}}(\varphi_f, t_f; \varphi_i, t_i) := \mathcal{A}^{\text{scalar}}(|\varphi_f\rangle, t_f; |\varphi_i\rangle, t_i) = \int_{\substack{\phi|_{t_i}=\varphi_i \\ \phi|_{t_f}=\varphi_f}} e^{iS[\phi]} \mathcal{D}\phi\,, \qquad (38.6)$$

where $S[\phi]$ is the classical action (the exact form is not important for the present discussion). Next, note that the field φ_i is a field on the hypersurface Σ_{t_i}, and φ_f is a field on the hypersurface Σ_{t_f}. These two hypersurfaces together form the boundary of the four-dimensional space–time region R, the region on which the field ϕ is defined. Let φ denote the combination of the fields φ_i, φ_f on the *full* boundary of R, denoted ∂R, which in this case is equal to $\Sigma_f \cup \Sigma_i$. The state $|\varphi_i\rangle$ can be thought of as living in a copy $\mathcal{H}_{\Sigma_{t_i}}$ of the Hilbert space associated with the surface Σ_{t_i}, and $|\varphi_f\rangle$ as living in a copy $\mathcal{H}_{\Sigma_{t_f}}$

of the Hilbert space of quantum states associated with the surface Σ_{t_f}. The full field φ on all of ∂R can then be thought of as labeling a state $|\varphi\rangle$ in a certain *combined* Hilbert space $\mathcal{H}_{\partial R}$ for the *full* boundary of R.

Let us define the Hilbert space $\mathcal{H}_{\partial R}$. Consider a given Hilbert space of quantum states $\mathcal{H}$. Often one thinks of quantum states $|\Psi\rangle \in \mathcal{H}$ as *column vectors* (*kets*). Their Hermitian conjugates, denoted $|\Psi\rangle^\dagger =: \langle\Psi|$, are then *row vectors* (*bras*). The inner product between two states Ψ, Φ can then be written as the matrix product of the row vector $\langle\Psi|$ with the column vector $|\Phi\rangle$, yielding a complex number

$$\langle\Psi|\Phi\rangle = \langle\Psi, \Phi\rangle$$

(whence the motivation for the notation $\langle\Psi|$ and $|\Phi\rangle$). The space of *row vectors* is called the space *dual* to $\mathcal{H}$, and is written $\mathcal{H}^*$. The Hilbert space $\mathcal{H}_{\partial R}$ for the *full* boundary $\partial R = \Sigma_{t_f} \cup \Sigma_{t_i}$, in terms of $\mathcal{H}_{\Sigma_{t_i}}$ and $\mathcal{H}_{\Sigma_{t_f}}$, is then defined to consist in formal sums of products of states in the dual $\mathcal{H}^*_{\Sigma_{t_f}}$ and in $\mathcal{H}_{\Sigma_{t_i}}$ (the product is denoted using the symbol $\otimes$). Mathematically, this is expressed by saying that $\mathcal{H}_{\partial R}$ is the *tensor product* of $\mathcal{H}^*_{\Sigma_{t_f}}$ with $\mathcal{H}_{\Sigma_{t_i}}$, and one writes $\mathcal{H}_{\partial R} := \mathcal{H}^*_{\Sigma_{t_f}} \otimes \mathcal{H}_{\Sigma_{t_i}}$. In terms of the initial and final field eigenstates $|\varphi_i\rangle \in \mathcal{H}_{\Sigma_{t_i}}, |\varphi_f\rangle \in \mathcal{H}_{\Sigma_{t_f}}$, the corresponding field eigenstate on the *full* boundary of R is given by $|\varphi\rangle := |\varphi_f\rangle^\dagger \otimes |\varphi_i\rangle = \langle\varphi_f| \otimes |\varphi_i\rangle \in \mathcal{H}_{\partial R}$. The Hilbert space $\mathcal{H}_{\partial R}$ on the full boundary of R is called the *boundary Hilbert space*, and $|\varphi\rangle$ is called a *boundary state.*

In terms of the label φ and *boundary* states, (38.6) becomes

$$\mathcal{A}^{\text{scalar}}(\varphi, R) \equiv \mathcal{A}^{\text{scalar}}(|\varphi\rangle, R) = \int_{\phi|_{\partial R}=\varphi} e^{iS[\phi]} \mathcal{D}\phi \,. \quad (38.7)$$

This expression has the benefit that it makes sense also when R is *any* space–time region, leading to a natural generalization of the path-integral formalism. This generalization is called the *general boundary* formulation of quantum mechanics, and is equivalent to the more standard formulations of quantum mechanics [38.2, 6, 7]. The interpretation of the path integral (38.7) is the direct generalization of the interpretation of the original path integral (38.6): it provides the probability amplitude of measuring the field ϕ to have the values φ on the boundary of the region R. The

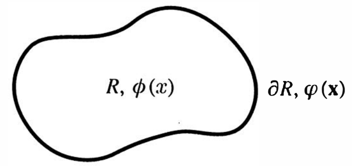


Fig. 38.3 The general boundary formulation of the path integral applies even when the space–time region R is compact

expression (38.7) applies when the boundary state is an eigenstate $|\varphi\rangle$ of the scalar field operator $\hat{\varphi}(\mathbf{x})$; from this one can deduce the amplitude $\mathcal{A}^{\text{scalar}}(\Psi, R)$ for *any* quantum boundary state Ψ in $\mathcal{H}_{\partial R}$. The general boundary formalism applies even, and in our case most importantly, when R is compact (Fig. 38.3). One advantage of this generalized formalism when R is chosen to be compact is that one can completely side step the issue of how the quantum state behaves as one approaches spatial infinity, an issue which should not matter for concrete applications anyway, because one never measures fields at infinity in actual experiments. Furthermore, the lack of an a priori fixed notion of which space–time regions may be used is more consistent with the spirit of background independence, which will be central in the case of quantum gravity.

38.1.3 The Case of Gravity: The *Problem of Time* and the Path Integral as Projector

Applying the above ideas to gravity involves unique subtleties. Specifically, in general relativity, when initial data surfaces are compact and without boundary (so that there are no boundary terms), the Hamiltonian H is constrained to be *zero.* In fact, the Hamiltonian can be expressed in terms of a Hamiltonian density $H = \int \mathcal{H}(\mathbf{x})\, d^3\mathbf{x}$, and this Hamiltonian density $\mathcal{H}(\mathbf{x})$ is constrained to be zero *at each point* $\mathbf{x}$. Because $\mathcal{H}(\mathbf{x})$ is constrained to be zero, it is called the Hamiltonian *constraint.* In the quantum theory, the Hamiltonian constraint dictates that states be eigenstates of the Hamiltonian constraint operator $\hat{\mathcal{H}}(\mathbf{x})$ with eigenvalue zero – that is, one requires that states be annihilated by the Hamiltonian constraint, $\hat{\mathcal{H}}(\mathbf{x})\Psi = 0$ and, hence, also by the Hamiltonian, $\hat{H}\Psi = 0$. By Schrödinger's equation (38.1), this implies the curious property

$$\frac{d\Psi}{dt} = \left(\frac{-i}{\hbar}\right)\hat{H}\Psi = 0 \,, \quad (38.8)$$

i. e., that the quantum state should not evolve in time. This fact is directly related to the background independence of general relativity: that there is no background time variable. Whereas in *classical* general relativity one can introduce an arbitrary time variable for convenience, in *quantum* general relativity, even introducing such a time for convenience is forbidden, or at least useless.

It is clear, therefore, that in quantum gravity one can not interpret the Feynman path integral in terms of time evolution, as was done in (38.4). In fact, the interpretation is different. Instead, in the interpretation of the path integral, the time-evolution map is replaced by a *projector* P onto solutions of $\hat{\mathcal{H}}(\mathbf{x})\Psi = 0$, the quantum Hamiltonian constraint [38.8–10]. Let us be concrete. In the case of gravity, the space–time field is the four-dimensional *metric*, denoted $g(x)$, which determines the lengths of, and angles between, vectors at each point x, which in turn determines geometrical lengths of curves, areas of surfaces, volumes of regions, etc. – that is, $g(x)$ determines the geometry of space–time. The *canonical* variables on a given instant-time hypersurface Σ_t are the *three*-dimensional metric $h(\mathbf{x})$ determining the three-dimensional geometry of Σ_t, and its conjugate momentum $\Pi(\mathbf{x})$, which determines the way Σ_t curves in the larger four-dimensional space–time and can be related to the time derivative of $h(\mathbf{x})$. Hence, in the quantum theory one has operators $\hat{h}(\mathbf{x})$ and $\hat{\Pi}(\mathbf{x})$, and simultaneous eigenstates $|h\rangle$ of the operators $\hat{h}(\mathbf{x})$. The states $|h\rangle$ and the projector P are then related to the Feynman path integral by

$$\langle h_\mathrm{f}, P\, h_\mathrm{i}\rangle = \int\limits_{\substack{g|_{\Sigma_{t_\mathrm{i}}} = h_\mathrm{i} \\ g|_{\Sigma_{t_\mathrm{f}}} = h_\mathrm{f}}} \mathrm{e}^{\mathrm{i}S[g]} \mathcal{D}g\,, \tag{38.9}$$

where $g|_\Sigma = h$ means that the geometry induced by g on Σ is equal to h.

Another way of stating this phenomenon is that (38.8) is simply a statement of *gauge invariance* of the wavefunction – time translations are coordinate transformations, and hence do not change the physical state, and so are gauge. At the same time, it is also a statement of the quantum version of the component $H = \int \mathcal{H}(\mathbf{x})\,\mathrm{d}^3\mathbf{x} = 0$ of the Hamiltonian constraint. In fact, in general, for every gauge symmetry in a system, there is a corresponding constraint and, as happens here, in the quantum theory, invariance under the gauge symmetry and satisfaction of the corresponding quantum constraint become one and the same thing. Constraints related to gauge in this way are called *first class* [38.11]. Not only is H a first-class constraint, but so are the infinity of individual Hamiltonian constraints $\mathcal{H}(\mathbf{x}) = 0$ for each point $\mathbf{x}$. In fact, all other fields which mediate forces in nature (electroweak and strong forces) also have first-class constraints and corresponding gauge symmetries. Quite generally, whenever a system has first-class constraints, the path integral projects onto solutions of the first-class constraints, so that the projection property seen in (38.9) is not unique to general relativity [38.8].

Exactly as in the case of the scalar field theory in the last subsection, (38.9) generalizes to an arbitrary space–time region. If $\mathcal{A}^{\mathrm{grav}}(\Psi, R)$ denotes the probability amplitude for a given quantum gravity state Ψ on the boundary of a given region R, h denotes a given three-dimensional metric on the boundary ∂R of R, and $|h\rangle$ the corresponding eigenstate in the boundary state space, we have

$$\mathcal{A}^{\mathrm{grav}}(h, R) := \mathcal{A}^{\mathrm{grav}}(|h\rangle, R) = \int\limits_{g|_{\partial R} = h} \mathrm{e}^{\mathrm{i}S[g]} \mathcal{D}g\,. \tag{38.10}$$

We close this section with a remark. In the case of a scalar field, there is a *background space–time geometry*, $\mathring{g}$, present and the action $S[\phi]$ depends on it: $S[\phi] = S[\phi, \mathring{g}]$. Because of this, $\mathcal{A}^{\mathrm{scalar}}(\Psi, R)$ in fact depends on the *size* and *shape* of the chosen region R, as determined by this background geometry. By contrast, in the case of quantum gravity, there is no background geometry, and so R has no nondynamically defined *shape* or *size*. In this case the boundary quantum state Ψ codes the information about geometry, which is now dynamical. If Ψ is sufficiently peaked on a classical geometry, then R again has a shape, but this shape is determined by Ψ, and not by any background geometry.

38.2 Spin-Foam Models of Quantum Gravity

38.2.1 Review of Spin-Network States and Their Meaning

It is now time to incorporate into the discussion what has been learned from the canonical quantization of gravity known as (canonical) loop quantum gravity. The Hilbert space of states in LQG is spanned by what are called *spin networks* (as discussed in Chap. 37). In this chapter, because it will be most useful later on, we review a form of spin network introduced by *Livine* and *Speziale* [38.12], which we will refer to as Livine–Speziale spin networks. (In the literature they are more commonly referred to as *Livine–Speziale coherent states*.) Each such spin network state is *peaked* on a particular three-dimensional, discrete, spatial geometry. We first review how each spin network is labeled, and then how these labels determine the corresponding geometry.

Each spin-network state is first labeled by a collection of curves in space which intersect each other at most at their end points. Such a collection of curves is called a *graph* and will be typically denoted γ (Fig. 38.4). Following the terminology of *Rovelli* [38.2], we call each curve in the graph a *link*, and each end point of a curve a *node*. Each link ℓ is labeled by a half integer spin $j_\ell = 0, \frac{1}{2}, 1, \frac{3}{2}, \ldots$ At each node ν, and for each link ℓ ending or beginning at ν, there is furthermore a unit, three-dimensional vector $n_{\nu\ell}$ (Fig. 38.5). We write $|\gamma, \{j_\ell, n_{\nu\ell}\}\rangle$ to denote such a spin network.

The labels $\gamma, \{j_\ell, n_{\nu\ell}\}$ determine the spatial geometry by determining areas of surfaces and volumes of regions. In determining these areas and volumes, an important role is played by the so-called *Planck length*, the unique combination, with dimension of length, of Newton's gravitational constant (G), Planck's constant divided by 2π ($\hbar$), and the speed of light (c). It is given by $\ell_{\text{Pl}} := \sqrt{G\hbar/c^3}$, which is approximately 1.616×10^{-35} m, or roughly 10 sextillionths of (or $10^{-20}\times$) the diameter of a proton. Given a surface S, in terms of the Planck length, its area as determined by a spin-network state with the labels $\gamma, \{j_\ell, n_{\nu\ell}\}$ is

$$A(S) = \sum_{\ell \text{ intersecting } S} 8\pi \ell_{\text{Pl}}^2 \beta \sqrt{j_\ell (j_\ell + 1)}\,, \qquad (38.11)$$

where β is a certain positive real number referred to as the Barbero–Immirzi parameter [38.13–16] (Fig. 38.6). Given a three-dimensional region R in space, its volume is

$$V(R) = \frac{(8\pi\beta)^{3/2}\ell_{\text{Pl}}^3}{4\sqrt{3}} \times \sum_{\nu \text{ nodes of } \gamma \text{ in } R} \sqrt{\left| \sum_{\ell,\ell',\ell'' \text{ at } \nu} j_\ell j_{\ell'} j_{\ell''}\, n_{\nu\ell} \cdot (n_{\nu\ell'} \times n_{\nu\ell''}) \right|}\,, \qquad (38.12)$$

where the sum over ℓ, ℓ', ℓ' is over all triples of links in γ starting or ending at the node ν.

38.2.2 Interpretation of Spin Networks in Terms of the Dual Complex

The extraction of information about geometry from the quantum labels j_ℓ, $n_{\nu\ell}$ can be systematized using what

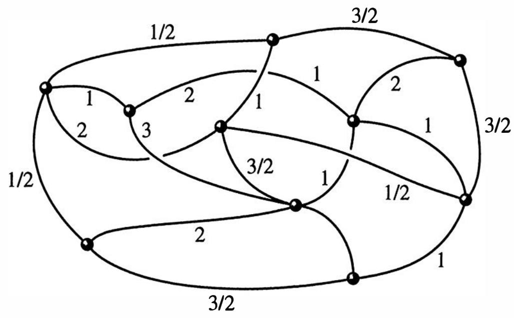


Fig. 38.4 Each spin-network state is labeled by a choice of graph, with *spins* labelling the links, and other *quantum numbers* labeling the nodes

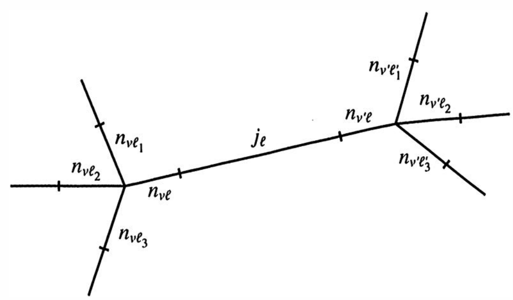


Fig. 38.5 Each link ℓ is labeled by a spin j_ℓ. For each node n, and each link ℓ incident at n, one has also a unit three-dimensional vector $n_{\nu\ell}$

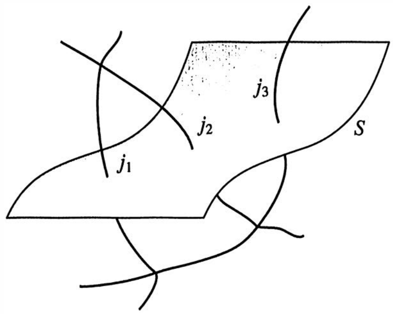


Fig. 38.6 Each spin-network link, with spin j_ℓ, intersecting a surface S contributes to the surface an area of $8\pi\ell_{\mathrm{Pl}}^2\beta\sqrt{j_\ell(j_\ell+1)}$, where ℓ_{Pl} is the Planck length

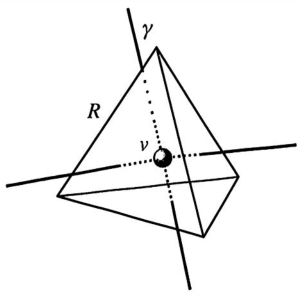


Fig. 38.7 A three-dimensional region $R=\nu^\star$ is said to be *dual* to a node ν of γ if it contains ν but no other node of γ

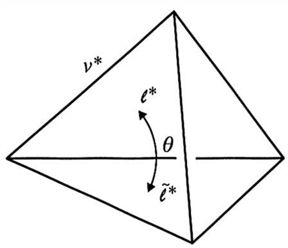


Fig. 38.8 The interior angle θ between the two faces $\ell^\star$ and $\tilde{\ell}^\star$ of the 3-cell $\nu^\star$, as determined by the LQG spin-network labels, is given by (38.14)

is called a *dual-cell complex.* For each link ℓ, a surface (a two-dimensional region) $S=\ell^\star$ is said to be *dual* to ℓ if it intersects ℓ at one point, but intersects no other link of γ. For each node ν, a three-dimensional region $R=\nu^\star$ is said to be *dual* to ν if it contains ν but no other node of γ (see Fig. 38.7).

If one chooses such a dual for each link and node in the graph γ, and if these are chosen such that they all *fit together* – that is, such that the boundary of each chosen three-dimensional region $\nu^\star$ consists entirely of chosen two-dimensional regions $\ell^\star$, then the set of all the chosen regions $\nu^\star$, $\ell^\star$ forms a *cell complex* which is said to be *dual to* γ, and which we denote by $\gamma^\star$. In this case, we refer to $\nu^\star$ and $\ell^\star$ as *cells* of $\gamma^\star$; more specifically, one uses the terms *3-cell* and *2-cell*, respectively, according to the dimension of the region. From (38.11), the spin j_ℓ on a link ℓ determines the area of the surface $\ell^\star$ dual to it by the formula

$$A(\ell^\star)=8\pi\ell_{\mathrm{Pl}}^2\beta\sqrt{j_\ell(j_\ell+1)}\,. \qquad (38.13)$$

From (38.12), the quantum labels $n_{\nu\ell}$ at a given node ν determine the volume of the region $\nu^\star$ dual to it via the formula

$$V(\nu^\star)=\frac{(8\pi\beta)^{3/2}\ell_{\mathrm{Pl}}^3}{4\sqrt{3}} \times\sqrt{\left|\sum_{\substack{\ell,\ell',\ell''\\ \text{at }\nu}} j_\ell j_{\ell'} j_{\ell''}\, n_{\nu\ell}\cdot(n_{\nu\ell'}\times n_{\nu\ell''})\right|}\,.$$

In addition to this, given a node ν and two links $\ell,\tilde{\ell}$ incident at it, one can ask what is the *angle* $\theta=\theta[\nu^\star,\ell^\star,\tilde{\ell}^\star]$ between the dual surfaces $\ell^\star,\ell'^\star$ within the dual region $\nu^\star$. In fact, it is given by the formula

$$\cos\left(\theta[\nu^\star,\ell^\star,\tilde{\ell}^\star]\right)=-n_{\nu\ell}\cdot n_{\nu\tilde{\ell}} \qquad (38.14)$$

(Fig. 38.8). These areas, volumes, and interior angles form the basic quantities from which the quantum geometry is constructed. We will go into more detail about this in Sect. 38.3.6.

There is of course a great deal of choice in the complex $\gamma^\star$ dual to γ. However, given γ, the *connectivity* of the parts of $\gamma^\star$ is uniquely determined – that is, which lower dimensional cells are on the boundary of each higher dimensional cell *is* uniquely determined. If ν is on the boundary of ℓ (meaning, in this case, an end point of ℓ), then $\ell^\star$ is on the boundary of $\nu^\star$. Another way of saying this, in mathematical terms, is that the *topology* of $\gamma^\star$ is unique, and it is in this sense that we can speak unambiguously of *the complex* $\gamma^\star$ *dual to* γ.

We have here discussed dual cells and dual-cell complexes in three dimensions. However, these ideas

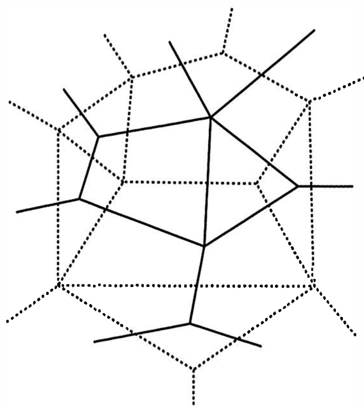


Fig. 38.9 Example of dual-cell complexes in two dimensions. The *solid line complex* and the *dotted line complex* are dual to each other

can be formulated in any dimension. If one is working in an N-dimensional space, and one has an M-dimensional surface S, a surface $S^\star$ is said to be *dual* to S if it has dimension $N-M$ and intersects S at exactly one point (Fig. 38.12). In the case of two dimensions, one can visualize the idea of a dual-cell complex with more completeness and ease. For the purpose of illustration, we include in Fig. 38.9 an example of a dual complex in two dimensions.

38.2.3 Histories of Spin Networks: Spin Foams

Histories of three-dimensional spin networks $|\gamma, \{j_\ell, n_{v\ell}\}\rangle$ become four-dimensional objects. The one-dimensional links of the graphs γ become two-dimensional *faces* f, and the zero-dimensional nodes

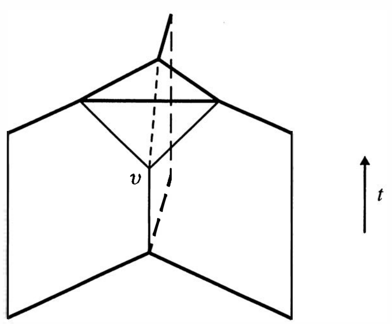


Fig. 38.10 A single node splits into three nodes, creating a spin-foam vertex

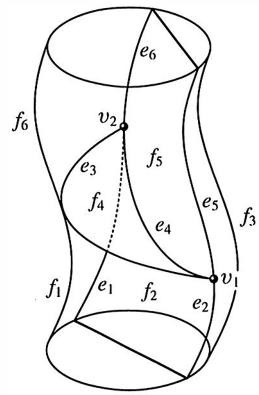


Fig. 38.11 A spin foam is a history of a spin network. It forms a *two-complex*, with the links of the spin network sweeping out faces, and the nodes of the spin network sweeping out edges. Each face f in the spin foam inherits the spin on the corresponding link, and each edge e in the spin foam inherits the set of unit three-dimensional vectors labeling the corresponding node. The face spins are now denoted j_f and the three- dimensional vectors are now denoted n_{ef}

of the graphs become one-dimensional *edges* e. Places in the history where a node splits into multiple nodes, or multiple nodes combine, are called *vertices* (Fig. 38.10). The set of all such faces, edges, and vertices of a given history together form the *spin-foam two-complex* of the history, which we usually denote $\mathcal{F}$ (Fig. 38.11). Each face f inherits the half-integer spin j_f labeling the link of which it is the history, and each edge e inherits the set of unit vectors associated with the node of which it is a history, one unit vector n_{ef} for each edge e and face f incident at e. The spin-foam two-complex $\mathcal{F}$, together with these labels, is referred to as a *spin foam*. Specifically, with this choice of labels, we will call it a *loop quantum gravity spin foam*. Each such spin foam represents, in a precise sense to be reviewed in Sect. 38.3.6, a *quantum space–time geometry*.

38.2.4 Spin-Foam Amplitudes

In order to specify the quantum dynamics, a *probability amplitude* must be specified for each spin foam – that is, a probability amplitude for each history of quantum gravity states, each *quantum space–time*. This amplitude should be, in an appropriate semiclassical limit, equal to (a possible real coefficient times) the usual

Feynman prescription of the exponential of i times the classical action, as reviewed in Sect. 38.1.1.

It turns out, from experience with simple theories in four space–time dimensions and gravity in three space–time dimensions [38.17, 18], that one expects this amplitude to be of the form

$$\mathcal{A}(\mathcal{F}, \{j_f, n_{ef}\}) = \left(\prod_{f\in\mathcal{F}} \mathcal{A}_f\right)\left(\prod_{e\in\mathcal{F}} \mathcal{A}_e\right)\left(\prod_{v\in\mathcal{F}} \mathcal{A}_v\right), \qquad (38.15)$$

where for each face f, edge e, and vertex v, $\mathcal{A}_f$, $\mathcal{A}_e$, and $\mathcal{A}_v$ are referred to as the *face*, *edge*, and *vertex* amplitudes, respectively. This form of the probability amplitude is called the *spin-foam ansatz*. Here, $\mathcal{A}_f$ is a function of the spin j_f alone, $\mathcal{A}_e$ is a function of the quantum labels associated with the edge e as well as with the faces incident at e, and $\mathcal{A}_v$ is a function of the quantum labels associated with the edges and faces incident at the vertex v. From experience with the above-mentioned simple models, $\mathcal{A}_f$ and $\mathcal{A}_e$ are expected to be real, and $\mathcal{A}_v$ complex. Thus, one expects the exponential of i times the action to arise almost entirely from the vertex amplitudes alone. It is for this reason that the vertex amplitude is usually considered the most important one. Furthermore, the vertices are where the spin network *changes* in the history, and hence where *interesting dynamics* is taking place. Thus, in a sense, it is not surprising that the vertex amplitude usually turns out to be the most important factor in the probability amplitude.

38.3 Deriving the Amplitude via a Simpler Theory

How should one determine the different factors $\mathcal{A}_f$, $\mathcal{A}_e$, and $\mathcal{A}_v$ appearing in the probability amplitude (38.15)? The strategy used by the spin-foam community is a bit indirect: we first construct the spin-foam amplitude for a very simple toy theory, called BF theory. The spin-foam dynamics of BF theory is very well understood. One then uses the fact that general relativity can be obtained from BF theory by imposing extra constraints, called *simplicity constraints*, an idea which traces back to the work of *Plebanski* [38.19]. The nontrivial task in constructing a spin-foam model is then reduced to the question of how these simplicity constraints should be imposed in the quantum theory.

We begin this section by reviewing a minimum necessary to understand what is BF theory, what are the simplicity constraints, and how Einstein's theory of gravity can be recovered from these. We then review the quantum mechanics of BF theory, and then discuss the version of the quantum simplicity constraints now predominant in the literature. (For the first method of imposing quantum simplicity which was previously predominant, and which laid the foundations for the modern method, see [38.20, 21].)

38.3.1 BF Theory and Gravity

BF theory is a theory with a maximal number of gauge symmetries. Recall that a *gauge symmetry* is a transformation that does not change the physical state of the system, but only changes the variables used to describe it. That is, the presence of gauge symmetries in a theory indicates a redundancy in the variables used to describe the system. The simplest example of a gauge symmetry is a transformation of the vector potential of the magnetic field: given a vector potential $\boldsymbol{A}$ and a function χ on space, the new vector potential

$$\tilde{\boldsymbol{A}} := \boldsymbol{A} + \nabla\chi$$

determines exactly the same magnetic field, and hence the same physical state of the system. In the case of general relativity, the gauge transformations are *space–time coordinate transformations*, reflecting the physical fact that space–time coordinates have no intrinsic meaning in the theory: space–time coordinates are only tools of convenience, used to aid in describing physical fields. The more gauge symmetries one has in a system, the less the variables of the theory contain real, physical information. BF theory has so many gauge symmetries that in fact the variables of the theory contain *no* local information. This is why BF theory is so simple and why the corresponding spin-foam quantum theory is so well understood. Such simple theories like BF theory which have no local physical degrees of freedom are referred to as *topological field theories*.

To introduce the basic variables of BF theory, we first recall a notation usually used for matrices: given a matrix M, one denotes its element in the i-th row and j-th column by M_{ij}. When one allows more than just two indices, one obtains a generalization of matrices, which

we shall call *arrays*. The basic variables of BF theory are two fields of arrays on space–time, denoted $\Sigma^{IJ}_{\mu\nu}(x)$ and $\omega^{IJ}_{\mu}(x)$, where the indices μ, ν, I, J take the values 0, 1, 2, 3. (There are more specific terms for these types of fields of arrays, which indicate certain transformation properties, but we have chosen to avoid these terms, because we wish to avoid talking about transformation properties which are not necessary for the discussion in this chapter.) In terms of these variables, the action for BF theory is given by

$$\begin{aligned} S_{\mathrm{BF}} &= \frac{1}{32\pi G} \\ &\times \int \epsilon^{\mu\nu\sigma\rho} \left(\frac{1}{2}\epsilon_{IJKL}\Sigma^{KL}_{\mu\nu} + \frac{1}{\beta}\eta_{IK}\eta_{JL}\Sigma^{KL}_{\mu\nu} \right) F^{IJ}_{\sigma\rho}\,\mathrm{d}^4x \\ &=: \frac{1}{2}\int \epsilon^{\mu\nu\sigma\rho} B_{\mu\nu IJ} F^{IJ}_{\sigma\rho}\,\mathrm{d}^4x\,, \end{aligned} \tag{38.16}$$

where ϵ_{IJKL}, $\epsilon^{\mu\nu\sigma\rho}$ both denote the *Levi-Civita array*, defined uniquely by the properties $\epsilon_{0123} = 1$ and that when any two indices of *IJKL* (respectively $\mu\nu\sigma\rho$) are interchanged, ϵ_{IJKL} (respectively $\epsilon^{\mu\nu\sigma\rho}$) changes by a minus sign – for example, $\epsilon_{IJKL} = -\epsilon_{JIKL}$. Furthermore, here and throughout the rest of this section we use the Einstein summation convention: when a given index appears twice in an expression, once up and once down, summation shall be implied over all possible values of the given index. In the final expression above, we have defined $B^{IJ}_{\mu\nu}$ to be the quantity in parentheses, and $F^{IJ}_{\sigma\rho}$ denotes the *field strength* (or *curvature*) of ω^{IJ}_{μ}, defined by

$$F^{IJ}_{\sigma\rho} := \frac{\partial \omega^{IJ}_{\rho}}{\partial x^{\sigma}} - \frac{\partial \omega^{IJ}_{\sigma}}{\partial x^{\rho}} + \eta_{KL}\left(\omega^{IK}_{\sigma}\omega^{LJ}_{\rho} - \omega^{IK}_{\rho}\omega^{LJ}_{\sigma}\right) .$$

(Note that in the spin-foam literature, sometimes $B^{IJ}_{\mu\nu}$ is defined to be only the first term of the expression in parentheses in (38.16).)

The simplicity constraint, in its simplest and most important sense, is just the requirement that there exist a matrix field $e^{I}_{\mu}(x)$ such that

$$\Sigma^{IJ}_{\mu\nu}(x) = \pm\left(e^{I}_{\mu}(x)e^{J}_{\nu}(x) - e^{J}_{\mu}(x)e^{I}_{\nu}(x)\right) \tag{38.17}$$

at each space–time point x. The simplicity constraint is thus a constraint on the field $\Sigma^{IJ}_{\mu\nu}(x)$; when $\Sigma^{IJ}_{\mu\nu}(x)$ satisfies this constraint, one says $\Sigma^{IJ}_{\mu\nu}(x)$ is *simple*. When $\Sigma^{IJ}_{\mu\nu}(x)$ is simple, the fields describing the system are just e^{I}_{μ} and ω^{IJ}_{μ}. What is important to know is that these fields in fact just describe a geometry for space–time. In terms of the metric tensor $g_{\mu\nu}$ used in much of this handbook, this geometry is just $g_{\mu\nu} = \eta_{IJ}e^{I}_{\mu}e^{J}_{\nu}$, where η_{IJ} is the diagonal matrix with $\eta_{11} = \eta_{22} = \eta_{33} = 1$, and $\eta_{00} = \pm 1$ depending on whether one is considering Euclidean $(+1)$ or Lorentzian (-1) gravity. e^{I}_{μ} is referred to as a *cotetrad*. The terms *Euclidean* and *Lorentzian* gravity are a bit misleading. In fact, there is only one Einstein theory of gravity describing the real world, and that is what we are calling here *Lorentzian* gravity. *Euclidean gravity* is a simplified model very closely related to, but in certain ways simpler than, Lorentzian gravity. It is often used for *practice* when investigating quantum gravity. Essentially, in Euclidean gravity one treats time as though it were just a fourth dimension of space. In this chapter, we consider spin-foam quantizations of both of these models of gravity.

38.3.2 Spin Foams of BF Theory

We have already described the spin foams arising from loop quantum gravity. We next describe spin foams for BF theory. These again arise as histories of labels of corresponding canonical quantum states, just as the loop quantum gravity spin foams of Sect. 38.2.3 arose as histories of labels of the canonical Livine–Speziale spin-network states of loop quantum gravity. As mentioned, there are two basic variables of BF theory, $\Sigma^{IJ}_{\mu\nu}$ and ω^{IJ}_{μ}. The restrictions of these fields to a given instant-time hypersurface are canonically conjugate, so that, to have a complete set of canonical states, it is sufficient to consider states peaked on $\Sigma^{IJ}_{\mu\nu}$ *or* ω^{IJ}_{μ}, but not both. The particular choice of canonical states we use for defining BF spin foams are peaked on the variable $\Sigma^{IJ}_{\mu\nu}$ and are closely related to the Livine–Speziale spin networks [38.22–25]. This choice will facilitate imposing the simplicity constraint, as well as be important for taking the semiclassical limit of the theory.

Exactly as in the case of the loop quantum gravity spin foams, each BF spin foam is first labeled by a *spin-foam two-complex*, with faces, edges, and vertices (as in Fig. 38.11). However, now the labels on the faces and edges are different. As mentioned at the end of the last section, in quantum gravity, often one considers first the simplified theory of Euclidean gravity for practice, before considering the actual Lorentzian gravity corresponding to reality. Spin foams are no exception. The spin-foam quantum labels for BF theory are different depending on whether one considers Euclidean or Lorentzian gravity. In the Euclidean case, each face f is labeled by two half integers $j_f^{+}, j_f^{-} = 0, \frac{1}{2}, 1, \frac{3}{2}, \ldots$ and,

Table 38.1 Labels on each face and on each edge bounding each face, for the BF spin-foam model used in the case of Euclidean and Lorentzian gravity, respectively

	For each face f	For each edge $e \in \partial f$
Euclidean	j_f^+, j_f^-	j_{ef}, n_{ef}
Lorentzian	p_f, k_f	j_{ef}, n_{ef}

for each edge e in the boundary of f, one has a further half integer j_{ef} and a unit three-dimensional vector n_{ef}. In the Lorentzian case, each face f is labeled by a real number p_f and a half integer k_f and, for each edge e in the boundary of f one again has a half integer j_{ef} and a unit three-dimensional vector n_{ef} (Table 38.1). For convenience, we let $\mathcal{L}$ denote the appropriate set of possible labels on the BF spin foam: $\{j_f^+, j_f^-, j_{ef}, n_{ef}\}$ in the Euclidean case, and $\{p_f, k_f, j_{ef}, n_{ef}\}$ in the Lorentzian case.

In terms of these labels, the amplitude for a single BF spin foam decomposes as in the spin-foam ansatz (38.15), with certain expressions for the corresponding face, edge, and vertex amplitudes $\mathcal{A}_f^{\mathrm{BF}}$, $\mathcal{A}_e^{\mathrm{BF}}$, and $\mathcal{A}_v^{\mathrm{BF}}$. What is important for this chapter is that the vertex amplitude can be expressed in terms of integrals over certain *groups*. Let $\mathcal{G}$ denote the space of all 4×4 matrices $G^I{}_J$ such that

$$G^K{}_I G^L{}_J \eta_{KL} = \eta_{IJ} ,$$

where η_{IJ} is again the diagonal four by four matrix with diagonal components $\eta_{11} = \eta_{22} = \eta_{33} = 1$ and $\eta_{00} = +1$ or -1 depending on whether one is considering Euclidean or Lorentzian gravity. For the case of Lorentzian gravity, a matrix $\mathcal{G}$ satisfies the above equation if and only if its action on a given set of four space–time coordinates is a *Lorentz transformation*; in this case $\mathcal{G}$ is called the *Lorentz group*. In the case of Euclidean gravity, $\mathcal{G}$ is the group of *four-dimensional Euclidean rotations*. (In fact, the group $\mathcal{G}$ is directly related to the labels on the faces of the BF spin foam: each pair (p_f, k_f) labels a unitary irreducible representation of the Lorentz group, and each pair (j_f^+, j_f^-) labels a unitary irreducible representation of the group of four-dimensional Euclidean rotations. This is similar to the way the angular momentum quantum number j in basic quantum mechanics labels irreducible representations of the spatial rotation group.) There is a way to define integrals over the group of matrices $\mathcal{G}$. The vertex amplitude $\mathcal{A}_v^{\mathrm{BF}}$ of BF theory can be expressed in terms of nested integrals over such matrices, one such integral for each edge e incident at the given vertex v

$$\mathcal{A}_v^{\mathrm{BF}}(\mathcal{L}) = \left(\prod_{e \text{ incident at } v} \int_{\mathcal{G}} \mathrm{d}G_{ve} \right) \tilde{\mathcal{A}}_v^{\mathrm{BF}}(\mathcal{L}, \{G_{ve}\}) . \tag{38.18}$$

One can think of the spin-foam two-complex $\mathcal{F}$, together with the labels $\mathcal{L}$ *and* the group matrices $\{G_{ve}\}$, as labelling a sort of *augmented* history, and $\tilde{\mathcal{A}}_v^{\mathrm{BF}}(\mathcal{L}, \{G_{ve}\})$ is the probability amplitude associated with this history. It is these augmented histories that will have a complete interpretation in terms of the classical variables of BF theory, as we will see in the next subsection. Beyond the above general form (38.18), the details of the vertex amplitude will not be needed in this chapter.

38.3.3 Dual-Cell Complex

To interpret the quantum labels for the BF spin foams in terms of classical BF theory, we use the same strategy as that used in Sect. 38.2.2 to interpret spin-network labels: we again use the notion of a *dual-cell complex*, except now in one dimension higher. In this subsection we explicitly spell out this duality in the four-dimensional case, lifting the duality presented in Sect. 38.2.2 from space to space–time.

Recall that each spin foam is first of all labeled by a spin-foam *two-complex* $\mathcal{F}$, consisting of vertices, edges, and faces, which fit together. For each vertex v in $\mathcal{F}$, a four-dimensional region $v^\star$ is said to be *dual* to v if it contains v and no other vertices of $\mathcal{F}$. For each edge e in $\mathcal{F}$, a three-dimensional hypersurface $e^\star$ is said to be *dual* to e if it intersects e in exactly one point, and intersects no other edges in $\mathcal{F}$. For each face f in $\mathcal{F}$, a two-dimensional surface $f^\star$ is said to be *dual* to f if it intersects f at one point, and intersects no other faces in $\mathcal{F}$. (For comparison with examples of dual cells when working in lower dimensions, see Fig. 38.12.) If one chooses such a dual for each vertex, edge, and face in $\mathcal{F}$, and if these are chosen such that they all *fit together* – that is, such that the boundary of each chosen four-dimensional region $v^\star$ consists entirely of chosen three-dimensional hypersurfaces $e^\star$, and the boundary of each three-dimensional hypersurface $e^\star$ consists entirely of chosen two-dimensional surfaces $f^\star$, then the set of all the chosen regions $v^\star$, $e^\star$, $f^\star$ forms a *cell complex* which is said to be *dual to* $\mathcal{F}$, and which we denote by $\mathcal{F}^\star$. In this case we refer to $v^\star$, $e^\star$, and $f^\star$ as *cells* of $\mathcal{F}^\star$; more specifically, one uses the terms

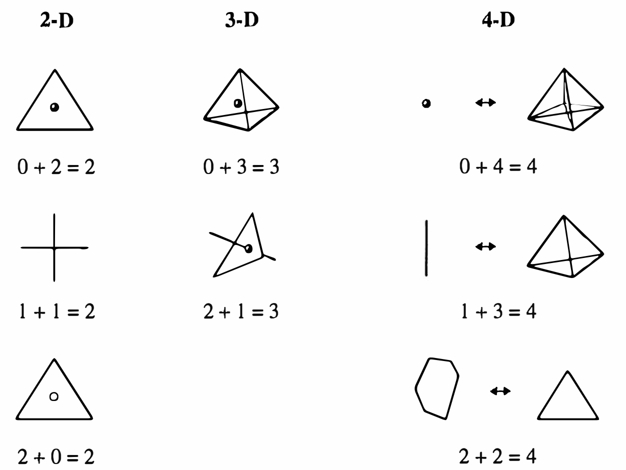


Fig. 38.12 Examples of dual cells in two, three, and four dimensions. For each pair of dual cells, the dimensionalities of the cells add up to the total dimensionality of the ambient space, and intersect in one point. In the four-dimensional case, the fact that dual cells intersect in one point can not be depicted

4-cell, *3-cell*, and *2-cell*, respectively, according to the dimension of the region. Once again, though there is a great deal of choice in such a complex $\mathcal{F}^\star$ dual to $\mathcal{F}$, the *connectivity* of the parts of $\mathcal{F}^\star$ is uniquely determined – that is, the *topology* of $\mathcal{F}^\star$ is unique.

38.3.4 Interpretation of the Labels

To interpret these labels, we arbitrarily fix a coordinate system x^μ in each 4-cell $v^\star$ such that, in this coordinate system, each 3-cell $e^\star$ and 2-cell $f^\star$ bounding $v^\star$ is *planar*. No physical quantities arising from the constructions that follow depend on this choice of coordinates in each $v^\star$. The classical field $\Sigma^{IJ}_{\mu\nu}$ corresponding to a given *augmented* BF spin foam $(\mathcal{F}, \mathcal{L}, G_{ve})$ is then constant in each 4-cell $v^\star$ (in the coordinates x^μ fixed in each $v^\star$). Let $(\Sigma_v)^{IJ}_{\mu\nu}$ denote the constant value taken by $\Sigma^{IJ}_{\mu\nu}(x)$ in the cell $v^\star$. The labels $\{j_{ef}, n_{ef}, G_{ve}\}$ are then related to $(\Sigma_v)^{IJ}_{\mu\nu}$ by

$$8\pi \ell_{\text{Pl}}^2 j_{ef} n^i_{ef} = \left((G_{ve})^0{}_L (G_{ve})^i{}_M + \frac{s}{\beta} (G_{ve})^j{}_L (G_{ve})^k{}_M \right) \int_f \Sigma_v^{LM} \tag{38.19}$$

for $(i,j,k) = (1,2,3), (3,1,2), (2,3,1)$, where $s = +1$ for Euclidean gravity and -1 for Lorentzian gravity, and where, recall, ℓ_{Pl} denotes the Planck length. The remaining labels, $\{j_f^+, j_f^-\}$ in the Euclidean case and $\{p_f, k_f\}$ in the Lorentzian case, are related to the classical field $\Sigma^{IJ}_{\mu\nu}(x)$ by

$$\eta_{IK}\eta_{JL}\left(\int_f \Sigma^{IJ}\right)\left(\int_f \Sigma^{KL}\right) = \begin{cases} C_{\text{E}}\left(\frac{(j_f^+)^2}{(\beta+1)^2} + \frac{(j_f^-)^2}{(\beta-1)^2}\right) & \text{in Eucl. case}\,, \\ C_{\text{L}}\left(k_f^2 - p_f^2 + \frac{4\beta}{1-\beta^2} kp\right) & \text{in Lor. case}\,, \end{cases} \tag{38.20}$$

$$\epsilon_{IJKL}\left(\int_f \Sigma^{IJ}\right)\left(\int_f \Sigma^{KL}\right) = \begin{cases} \tilde{C}_{\text{E}}\left(\frac{(j_f^+)^2}{(\beta+1)^2} - \frac{(j_f^-)^2}{(\beta-1)^2}\right) & \text{in Eucl. case}\,, \\ \tilde{C}_{\text{L}}\left(k_f + \frac{1}{\beta} p_f\right)\left(k_f - \beta p_f\right) & \text{in Lor. case}\,, \end{cases} \tag{38.21}$$

where $C_{\text{E}}, C_{\text{L}}, \tilde{C}_{\text{E}}, \tilde{C}_{\text{L}}$ are each a certain combination of the Planck length ℓ_{Pl}, β, and numerical factors. The integral $\int_S \Sigma^{IJ}$ of $\Sigma^{IJ}_{\mu\nu}(x)$ over a surface S appearing in the above equations is the standard *differential form* integral, defined by

$$\int_S \Sigma^{IJ} := \int \Sigma^{IJ}_{\mu\nu} \frac{\partial \tau^\mu}{\partial u} \frac{\partial \tau^\nu}{\partial v} \, \mathrm{d}u \, \mathrm{d}v\,, \tag{38.22}$$

where (u, v) are any choice of coordinates on S, and $x^\mu = \tau^\mu(u, v)$ is the four-dimensional position of the point on S with surface coordinates (u, v). (The result of the integral (38.22) is independent of the choice of (u, v) and hence of $\tau^\mu(u, v)$.)

38.3.5 Simplicity and the LQG Spin-Foam Model

Recall the general strategy we are taking: one starts from the probability amplitude $\mathcal{A}^{\text{BF}}(\mathcal{F}, \mathcal{L})$ for BF theory, and then restricts consideration to the case in which the BF spin foam $(\mathcal{F}, \mathcal{L})$ satisfies some quantum version of the simplicity constraint (38.17). Just as the classical simplicity constraint is sufficient to recover classical gravity from BF theory, so too one expects an appropriate quantum simplicity constraint to recover quantum gravity from quantum BF theory.

Different ways of imposing simplicity quantum mechanically then lead to different spin-foam models of

gravity. We will present here only the most recent, commonly used way of imposing quantum simplicity, which leads to the so-called *LQG spin-foam model.* This model is also variously referred to as the *EPRL*, *EPRL-FK*, or *EPRL-KKL* model, after key authors who contributed to its development [38.26–28] – namely the present author, Pereira, Rovelli, Livine, Freidel, Krasnov, Kamiński, Kisielowski, and Lewandowski. At the center of this strategy of imposing simplicity is the so-called *linear simplicity constraint*: the condition that

$$(G_{ve})^0{}_I \int_f \Sigma^{IJ} = 0 \tag{38.23}$$

for all f, e, and v incident on one another. It is called *linear* because it is linear in the field $\Sigma^{IJ}_{\mu\nu}(x)$. One can show that it implies that $\Sigma^{IJ}_{\mu\nu}$ takes one of the following three forms [38.29, 30]:

$$(\pm) \quad \Sigma^{IJ}_{\mu\nu}(x) = \pm\left(e^I_\mu(x)e^J_\nu(x) - e^J_\mu(x)e^I_\nu(x)\right) \text{ for some } e^I_\mu(x)\,, \tag{38.24}$$

$$(\text{deg}) \quad \Sigma^{IJ}_{\mu\nu}(x) \text{ is degenerate, that is, } \epsilon_{IJKL}\epsilon^{\mu\nu\rho\sigma}\Sigma^{IJ}_{\mu\nu}\Sigma^{KL}_{\rho\sigma} = 0\,. \tag{38.25}$$

Each of these constitutes a different sector of solutions to (38.23); we have chosen the symbols (+), (−), and (deg) to denote these sectors. Notice that only sectors (+) and (−) yield a field Σ of the form (38.17) required to obtain gravity. In fact, as we will see later in the section on the semiclassical limit, the existence of the last, degenerate, sector will cause problems, and we will mention one way to solve this problem. (However, it should be noted that the above three sectors of linear simplicity are already an improvement over the prior version of the simplicity constraint used in the literature [38.20–23, 31], which had five sectors.)

We just have discussed the *classical* implications of the linear simplicity constraint (38.23); however, it is the *quantum* implications for the BF spin foams that will yield us our quantum theory of gravity. From (38.19)–(38.21), one can deduce the consequences of linear simplicity (38.23) for the quantum numbers labeling the BF spin foams. In the Euclidean case, these are precisely

$$j_f^{\pm} = \frac{1}{2}|1 \pm \beta| j_{ef}$$

and, in the Lorentzian case,

$$p_f = \beta j_{ef} \text{ and } k_f = j_{ef}\,,$$

both remarkably simple forms. These are the quantum simplicity constraints at the heart of the LQG spin-foam model of gravity. After one imposes these constraints, one can ask: what free spin-foam labels are left? In the Euclidean case, one starts out with the BF spin-foam labels $\{j_f^+, j_f^-, j_{ef}, n_{ef}\}$; the above quantum simplicity constraint uniquely determines the labels $j_f^{\pm}$ in terms of j_{ef}, and furthermore forces that, for each f, all the spins j_{ef} to be equal, whence we can write simply j_f. Thus, the remaining free labels are $\{j_f, n_{ef}\}$. The same is true in the Lorentzian case: there one starts with the labels $\{p_f, k_f, j_{ef}, n_{ef}\}$, simplicity determines p_f and k_f in terms of j_{ef}, and all the spins j_{ef} for a given face f are equal, whence we may write j_f, and again the remaining free labels are $\{j_f, n_{ef}\}$. The key thing to note here is that *in both cases, the remaining free labels are exactly the same as the labels on the LQG spin foams introduced earlier.* Thus, just as *classically* the simplicity constraint reduces BF theory to gravity, so the quantum simplicity constraint reduces BF spin foams to LQG spin foams. That this key classical property is reproduced quantum mechanically is one of the principal successes of the linear simplicity constraint as imposed in the LQG spin-foam model, and is what allows the LQG spin-foam model to provide a dynamics for LQG, making it the first, and thus far only, spin-foam model to do so. For other, more subtle, but no less interesting, arguments for this model, we refer the reader to the original papers [38.26–28, 32].

38.3.6 Interpretation of LQG Spin-Foam Quantum Numbers: Quantum Space–Time Geometry

LQG spin foams describe the *gravitational field*, and hence the *geometry of space–time.* We here take time to explain how the labels of a LQG spin foam determine a *discrete space–time geometry*. This is important not only for understanding the meaning of the LQG spin-foam labels, but will be central in looking at the semiclassical limit of the resulting spin-foam quantum theory – that is, the limit in which quantum mechanics is *turned off*, and in which one should recover classical general relativity.

This section builds on Sects. 38.2.1 and 38.2.2 on the discrete spatial geometry determined by spin networks. Just as in classical general relativity, where spatial geometry fits into the larger space–time geometry, so too the quantum spatial geometry of spin networks fits consistently into the larger quantum space–time geometry of spin foams.

Interpretation of Quantum Numbers and Uniqueness of the Space–Time Geometry

The quantum numbers $\{j_f, n_{ef}\}$ determine a discrete space–time geometry by determining uniquely the geometry of *each 3-cell* $e^\star$ in the dual complex $\mathcal{F}^\star$. Specifically, for each 3-cell $e^\star$, the area of each face $f^\star$ of $e^\star$ is equal to $8\pi\ell_{\text{Pl}}^2\beta\sqrt{j_f(j_f+1)}$, and the interior angle $\theta[e^\star,f^\star,\tilde{f}^\star]$ between each pair of faces $f^\star,\tilde{f}^\star$ within $e^\star$ is given by the equation

$$\cos\left(\theta[e^\star,f^\star,\tilde{f}^\star]\right) = -n_{ef}\cdot n_{e\tilde{f}}\,.$$

These areas and angles are precisely the same areas and angles used to interpret the LQG spin-network labels in Sect. 38.2.2, except now in a four-dimensional context. Using a theorem by *Minkowski* [38.33], one can show [38.34] that these areas and angles are sufficient to uniquely determine a flat geometry within each 3-cell $e^\star$. Because one is now working in four dimensions, each 3-cell is now part of the boundary of two *4-cells*. The geometry of all the 3-cells $e^\star$ bounding each 4-cell $v^\star$ is sufficient to determine a flat geometry within each 4-cell [38.24, 35]. By determining the geometry within each 4-cell, one determines a geometry of the entire space–time. This geometry is *piecewise flat*: within each 4-cell it is flat, but the resulting overall geometry of the larger space–time certainly need not be flat, and indeed can approximate any desired space–time geometry arbitrarily well. (See Fig. 38.13 for a depiction of this phenomenon in two dimensions.) This is the discrete, quantum space–time geometry determined by a loop quantum gravity spin foam $(\mathcal{F},\{j_f,n_{ef}\})$.

Existence

The above discussion explains how the quantum numbers $\{j_f,n_{ef}\}$ of a LQG spin foam are sufficient to uniquely determine a piecewise-flat space–time geometry, *assuming* there *exists* a space–time geometry compatible with the given spin-foam data. This will not always be the case: there are constraints on the spin-foam data. Specifically, in order for a compatible space–time geometry to exist, two constraints must be satisfied [38.24, 25]: the *closure* and *gluing* constraints [38.36, 37]. The closure constraint requires that, in each dual 3-cell $e^\star$, one has

$$\sum_{f \text{ incident at } e} j_f n_{ef}^i = 0\,.$$

By the same theorem of Minkowski cited earlier, this constraint is sufficient to ensure that a consistent geometry for the *3-cell* $e^\star$ can be reconstructed. The second constraint, the *gluing* constraint, is the requirement that all of these 3-cell geometries *fit* together consistently – that is, when two 3-cells share a face, they should have the same area and shape. When this is true, then the *4-cell* geometries will also exist and fit together, yielding a full piecewise-flat space–time geometry consistent with the given spin-foam data.

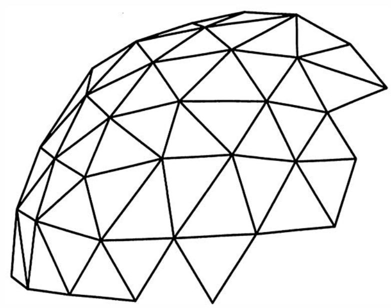


Fig. 38.13 Illustration in two dimensions: even though each 2-cell is flat, when many are glued together, the resulting two-dimensional cell complex need not be flat, and in fact can approximate any curved geometry

38.3.7 The Loop-Quantum-Gravity Spin-Foam Amplitude

We are now ready to implement the last step of the strategy to define a spin-foam model of quantum gravity. We have already described BF theory and its quantum histories in Sect. 38.3.1–38.3.2, and have introduced a way of imposing the simplicity constraint quantum mechanically in Sect. 38.3.5. Just as the set of classical BF fields satisfying classical simplicity coincides with the fields describing classical gravity, so too we have seen that the set of BF spin foams satisfying quantum simplicity as presented above is in 1–1 correspondence with loop quantum gravity spin foams. Let $\mathcal{I}$ denote the 1–1 map from LQG spin-foam labels on a given two-complex $\mathcal{F}$ to BF spin-foam labels satisfying simplicity on the same $\mathcal{F}$. The last step of the strategy is to restrict the spin-foam amplitude $\mathcal{A}^{\text{BF}}(\mathcal{F},\mathcal{L})$ of BF theory to spin foams satisfying simplicity. This leads one to assign the following probability amplitude to each LQG spin foam $(\mathcal{F},\{j_f,n_{ef}\})$

$$\mathcal{A}^{\text{LQG}}(\mathcal{F},\{j_f,n_{ef}\}) := \mathcal{A}^{\text{BF}}(\mathcal{F},\mathcal{I}(\{j_f,n_{ef}\}))\,. \quad (38.26)$$

This is the LQG spin-foam amplitude, defining the LQG spin-foam model of quantum gravity. It exists in both Euclidean and Lorentzian versions, depending on which BF theory one starts with. $\mathcal{A}^{\mathrm{LQG}}(\mathcal{F}, \{j_f, n_{ef}\})$ gives the probability amplitude for the single quantum space–time history $(\mathcal{F}, \{j_f, n_{ef}\})$, the geometrical meaning of which has been explained in the previous section. This is the principal spin-foam amplitude that will be used in the rest of this chapter.

Note that while the BF spin-foam amplitude is a well-established result of an exactly soluble theory, the LQG spin-foam amplitude (38.26) must be considered a proposal due to the nontrivial decision involved in the way the simplicity constraint is imposed. Nevertheless, the particular way of imposing the simplicity constraint presented above has compelling properties [38.27, 38], especially the exact reduction of BF spin-foam labels to those of LQG, which no other strategy thus far has.

38.4 Regge Action and the Semiclassical Limit

We turn attention now to the classical limit of the LQG spin-foam model. Recall from Sect. 38.1.1 that the classical limit is defined as the limit in which appropriate combinations of physical quantities become large compared with Planck's constant. In the case of gravity, the relevant physical quantities are *geometrical* and have dimensions of some power of length. As mentioned earlier in this chapter, there is a unique combination of Newton's gravitational constant, Planck's constant divided by 2π, and the speed of light, with dimension of length: the Planck length, $\ell_{\mathrm{Pl}} = \sqrt{G\hbar/c^3}$. The classical limit of a quantum theory of gravity arises when geometrical quantities become large compared to the corresponding power of the Planck length. In this limit, quantum theory can be neglected and, in order for the theory to remain compatible with the many successful experimental and observational tests of general relativity, it is necessary for the theory, in this limit, to become general relativity.

Recall from Sect. 38.1.1 that, when a quantum theory is formulated in terms of a path integral, the form e^{iS} of the amplitude for individual histories is important not only to have equivalence with the canonical quantum dynamics, but also important for ensuring the correct classical limit of the theory.

This leads us to ask: is the LQG spin-foam amplitude, derived above, equal to e^{iS}, with S an appropriate action for gravity? Except for one subtle point to be discussed further at the end of this subsection, this question has been answered in the affirmative in the limit in which geometrical quantities are large compared to the Planck scale, and in the special case in which the dual-cell complex $\mathcal{F}^\star$ consists of cells of the simplest type, called *simplices*. The limit in which geometrical quantities are large compared to the Planck scale is of course just the classical limit. However, because the probability amplitude for individual histories is a fundamentally quantum mechanical object with no classical analogue, one usually instead refers to this as the *semiclassical limit* of the amplitude. When all cells of $\mathcal{F}^\star$ are simplices, $\mathcal{F}^\star$ is called a *simplicial complex*. The possible piecewise-flat geometries on such a complex are called *Regge geometries*. Before stating the semiclassical limit of the LQG spin-foam amplitude, we explain in more detail simplicial complexes and Regge geometries.

38.4.1 Regge Geometries

Until this point we have spoken of general *cells* in the dual-cell complex $\mathcal{F}^\star$. In each dimension n, there is a certain type of simplest possible cell called a *simplex*, plural *simplices*. When one wishes to specify the dimension n of a simplex, one uses the term *n-simplex*. 0-simplices are *points*, 1-simplices are *line segments*, 2-simplices are *triangles*, and 3-simplices are *tetrahedra*. In four dimensions, there is no common term for the simplest possible cell; it is therefore simply called a 4-simplex (Fig. 38.14).

Recall from Sect. 38.3.6 how each LQG spin foam $(\mathcal{F}, \{j_f, n_{ef}\})$ determines a space–time geometry which is *flat* in each dual 4-cell $v^\star$ – that is, a *piecewise-flat space–time geometry*. When we furthermore require that all of the dual 4-cells be *simplicial*, the resulting geometry is called a *Regge geometry*. Regge geometries were first introduced by *Regge* [38.39] and are well studied in the literature. Usually Regge geometries are specified by giving the lengths of all 1-simplices, as this information is equivalent to specifying a piecewise-flat geometry, as noted in Regge's original paper. We furthermore note that the Regge geometries determined by spin foams are slightly more restricted than what is usual for Regge calculus, in that the areas of triangles are restricted to belong to the canonical area spectrum given in (38.13).

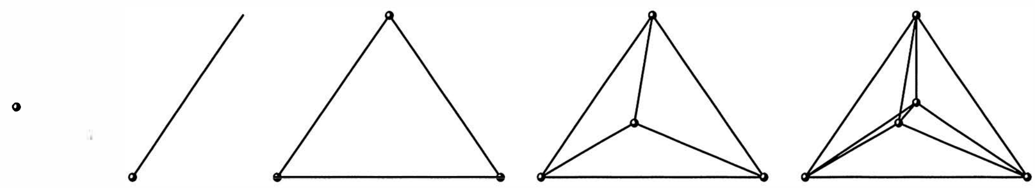


Fig. 38.14 A 0-simplex, a 1-simplex, a 2-simplex, a 3-simplex, and a 4-simplex, projected into a three-dimensional plane for visualization

When one takes the standard action for general relativity, the Einstein–Hilbert action, and evaluates it on Regge geometries, one obtains the *Regge action* [38.40]. Thus, in order to ensure that our spin-foam model of gravity has the correct classical limit (namely, general relativity), one would like the amplitude for such a spin foam to be

$$\mathcal{A}(\mathcal{F}, \{j_f, n_{ef}\}) = (\text{positive real number}) \exp(\mathrm{i}S_\mathrm{R}) \ ,$$

where S_R is the Regge action.

38.4.2 Semiclassical Limit

With the above background, we are ready to state the result on the semiclassical limit of the LQG spin-foam amplitude. The LQG spin-foam amplitude follows the spin-foam ansatz (38.15), so that it decomposes into face, edge, and vertex amplitudes

$$\begin{aligned}&\mathcal{A}^{\mathrm{LQG}}(\mathcal{F}, \{j_f, n_{ef}\})\\ &= \left(\prod_{f\in\mathcal{F}} \mathcal{A}^{\mathrm{LQG}}{}_f\right)\left(\prod_{e\in\mathcal{F}} \mathcal{A}^{\mathrm{LQG}}{}_e\right)\left(\prod_{v\in\mathcal{F}} \mathcal{A}^{\mathrm{LQG}}{}_v\right) .\end{aligned} \tag{38.27}$$

Recall that every complex number A can be decomposed as $A = |A|\mathrm{e}^{i\theta}$, where θ is the *phase*. Following the argument of Sect. 38.1.1, in order to obtain the correct classical limit of the quantum theory, it is sufficient for the *phase* of the amplitude to become the classical action. The face and edge amplitudes are real. Hence, they can contribute at most an integer multiple of π to the phase of the full amplitude. The interesting contribution to the phase of the amplitude will thus be from the vertex amplitudes, and so one focuses on the semiclassical limit of primarily the vertex amplitudes.

Each vertex amplitude $\mathcal{A}^{\mathrm{LQG}}{}_v$ can literally be understood as the spin-foam amplitude for a single 4-cell $v^\star$. As mentioned above, the semiclassical limit has thus far only been carried out for the case in which each 4-cell is a *4-simplex*, and thus we restrict consideration to the case in which $v^\star$ is a 4-simplex. In this case, the vertex amplitude $\mathcal{A}^{\mathrm{LQG}}{}_v$ depends on the 10 spins j_f on the 10 faces incident at v, on the 20 vectors n_{ef} labeling the five edges e incident at v, and on the four faces f incident at each of these five edges. To emphasize this dependence, we write $\mathcal{A}^{\mathrm{LQG}}{}_v(\{j_f, n_{ef}\})$, where it is understood that there is only dependence on these spins and vectors.

Recall that the semiclassical is the limit in which geometric quantities become large compared to the Planck scale. All geometric quantities determined by the labels $\{j_f, n_{ef}\}$ scale directly with the spins j_f, so that an easy way to take the semiclassical limit is to rescale all of these spins by some common parameter λ, and then take the limit in which λ becomes large. Thus, concretely, to look at the semiclassical limit of the vertex amplitude, one looks at the limit of $\mathcal{A}^{\mathrm{LQG}}{}_v(\{\lambda j_f, n_{ef}\})$ as λ becomes large.

The form of the semiclassical limit of $\mathcal{A}^{\mathrm{LQG}}{}_v(\{\lambda j_f, n_{ef}\})$ is different depending on the type of geometry, or lack thereof, determined by the labels $\{j_f, n_{ef}\}$. The form of the semiclassical limit of the Euclidean version of the vertex amplitude falls into three different cases, whereas that of the Lorentzian version is more subtle and falls into four different cases. Because the Euclidean case is sufficient to demonstrate the important issues, and is simpler, we restrict the following presentation to the Euclidean case. In Sect. 38.3.6 we reviewed how, if the spin-foam labels $\{j_f, n_{ef}\}$ satisfy what are called the closure and the gluing constraints, then they uniquely determine a (possibly degenerate) flat geometry for the 4-cell $v^\star$ (which in this case is a 4-simplex). If the tetrahedra on the boundary of $v^\star$, as determined by this geometry, all have nonzero volume, we say that the labels determine a *nondegenerate boundary geometry* of $v^\star$. If any of the tetrahedra have zero volume, we say that the labels $\{j_f, n_{ef}\}$ determine a *degenerate boundary geometry*. If the labels do not satisfy the

closure and gluing constraints, we say that they are *nongeometric*. We give below the semiclassical limit, which we denote by the symbol $\sim$, of the Euclidean version of $\mathcal{A}^{\text{LQG}}{}_v$ in each these three different cases:

1. For labels determining a nondegenerate boundary geometry

$$\mathcal{A}^{\text{LQG}}{}_v(\{\lambda j_f, n_{ef}\}) \sim \lambda^{-12}\left(C_1 e^{iS_R} + C_2 e^{-iS_R} + C_3 e^{\frac{i}{\beta}S_R} + C_4 e^{-\frac{i}{\beta}S_R}\right). \tag{38.28}$$

2. For labels determining a degenerate boundary geometry

$$\mathcal{A}^{\text{LQG}}{}_v(\{\lambda j_f, n_{ef}\}) \sim \lambda^{-12} C. \tag{38.29}$$

3. For nongeometric labels, the probability amplitude $\mathcal{A}^{\text{LQG}}{}_v(\{\lambda j_f, n_{ef}\})$ decays exponentially with λ, that is, faster than any inverse power of λ, so that such labels are *suppressed* by the vertex amplitude.

In the above formulae, C_1, C_2, C_3, C_4, and C are independent of λ. Note that the only labels not suppressed are the ones that actually correspond to piecewise-flat (possibly degenerate) space–time geometries. In addition to this, the fact that the exponential of i times various multiples of the action appears in the semiclassical limit of the vertex amplitude is encouraging. However, this is not yet sufficient to ensure the correct classical limit: not only are some unphysical, degenerate geometries not suppressed, but even for the nondegenerate geometries, the asymptotic amplitude (38.28) is not yet the Feynman amplitude. The Feynman amplitude would consist of only the first term in (38.28). There is a reason for the nonsuppression of the degenerate configurations in (38.29) as well as the extra terms in (38.28); in a moment, we will remark on this reason, as well as mention a solution to the problem. That these extra terms spoil the classical limit of the theory can be seen by looking at spin foams on triangulations with more than one 4-simplex. In this case, even if we assume that the geometries of all 4-simplices are nondegenerate, one still has the four terms in (38.28) *for each 4-simplex.* When these four terms are substituted into the expression (38.27) for the full amplitude, one obtains cross-terms. Each of these cross-terms is equal to the exponential of a sum of terms, one for each 4-simplex, equal to the Regge action for that 4-simplex times differing coefficients, yielding what can be called a *generalized Regge action* [38.30, 41, 42]. The extrema of this *generalized Regge action* are *not* the Regge equations of motion and hence *not* those of general relativity, so that general relativity fails to be recovered in the classical limit.

As shown in the recent work [38.22, 23], the extra terms causing this problem are due precisely to the presence of the multiple sectors of solutions to the simplicity constraint presented in Sect. 38.3.1, as well as the presence of different *orientations* as dynamically determined by the cotetrad field e^I_μ. Once these sectors and orientations are properly handled [38.29, 30], one arrives at what is called the *proper loop quantum gravity vertex amplitude.* Its semiclassical limit includes only the single term consisting of the exponential of i times the classical action

$$\mathcal{A}^{(+)}_v(\{\lambda j_f, n_{ef}\}) \sim \lambda^{-12} C_1 e^{iS_R},$$

thereby solving the above problem and giving reason to believe that the resulting spin-foam model will yield a correct classical limit.

38.5 Two-Point Correlation Function from Spin Foams

In this section, we review a calculation in spin foams which has played an important role in the development of the field: the calculation of the two-point correlation function of quantum gravity. The two-point correlation function of a quantum field theory is the simplest quantity one can calculate which directly probes the *nonclassical-ness* of the theory and thus provides one with a genuinely quantum mechanical prediction. In order to set the stage for this calculation, we begin with a technical discussion of how one sums over spin foams.

38.5.1 The Complete Sum over Spin Foams

Let us first recall how the spin-foam amplitude discussed in the last few sections fits into the overall calculation scheme. As discussed in Sect. 38.1, the amplitude for a given gravitational history is used to calculate the *probability amplitude* for a *canonical quantum state* on the boundary of a given space–time region. In the case where this canonical quantum state is an eigenstate $|h\rangle$ of spatial geometry on the boundary of some space–

time region R, the probability amplitude takes the form (38.10)

$$\mathcal{A}^{\text{grav}}(|h\rangle, R) = \int_{g|_{\partial R}=h} e^{iS[g]} \mathcal{D}g \,. \quad (38.30)$$

Spin foams provide a way to make the above formal prescription concrete, by using the lessons of loop quantum gravity. Loop quantum gravity tells us that the correct eigenstates of spatial geometry on ∂R are the *spin network states* $|\gamma, \{j_\ell, n_{v\ell}\}\rangle$, labeled by a graph γ on the boundary of R, spins j_ℓ, and unit 3-vectors $n_{v\ell}$ as in Sect. 38.2.1. The integral over continuum geometries is then replaced by the sum over discrete space–time geometries represented by spin foams. The formal expression (38.30) is then replaced by the concrete spin-foam expression

$$\mathcal{A}^{\text{LQG}}(|\gamma, \{j_\ell, n_{v\ell}\}\rangle, R) := \sum_{\substack{\mathcal{F} \text{ such that} \\ \mathcal{F}\cap\partial R=\gamma}} \sum_{\substack{\{j_f, n_{ef}\} \text{ such that} \\ \{j_f, n_{ef}\}|_{\partial R}=\{j_\ell, n_{v\ell}\}}} \mathcal{A}^{\text{LQG}}(\mathcal{F}, \{j_f, n_{ef}\}) \,.$$

Before using this expression, there is still one more issue that must be addressed. Two different types of apparently infinite sums appear in the above expression: (1) the sum over possible *two-complexes* $\mathcal{F}$ and (2) the sum over possible *labels* $\{j_f, n_{ef}\}$ on the two-complex. There is an infinite number of two-complexes, giving rise to the first potential source of infinity, and, for each two-complex, there are an infinite number of possible ways to label it, giving rise to the second potential source of infinity.

To aid in addressing the first of these potential infinities, it is useful to separate the sum over two-complexes into first a sum over *numbers of vertices N*, and then a sum over two-complexes with N vertices

$$\mathcal{A}^{\text{LQG}}(|\gamma, \{j_\ell, n_{v\ell}\}\rangle, R) := \sum_{N=0}^{\infty} \sum_{\substack{\mathcal{F} \text{ with } N \text{ vertices,} \\ \text{such that} \\ \mathcal{F}\cap\partial R=\gamma}} \sum_{\substack{\{j_f, n_{ef}\} \text{ such that} \\ \{j_f, n_{ef}\}|_{\partial R}=\{j_\ell, n_{v\ell}\}}} \mathcal{A}^{\text{LQG}}(\mathcal{F}, \{j_f, n_{ef}\}) \,.$$

For each N there are only a finite number of two-complexes, so the the potential infinity resides only in the sum over N. To handle this, one usually introduces a small, positive real number λ, raised to the power N

$$\mathcal{A}^{\text{LQG}}(|\gamma, \{j_\ell, n_{v\ell}\}\rangle, R) := \sum_{N=0}^{\infty} \lambda^N \sum_{\substack{\mathcal{F} \text{ with } N \text{ vertices,} \\ \text{such that} \\ \mathcal{F}\cap\partial R=\gamma}} \sum_{\substack{\{j_f, n_{ef}\} \text{ such that} \\ \{j_f, n_{ef}\}|_{\partial R}=\{j_\ell, n_{v\ell}\}}} \mathcal{A}^{\text{LQG}}(\mathcal{F}, \{j_f, n_{ef}\}) \,. \quad (38.31)$$

This has the effect of making each consecutive term in the sum over N smaller, and so ensuring convergence. The insertion of the power of λ not only brings this potential infinity under control, it also allows the resulting spin-foam theory to be recast in terms of something called a *group field theory* [38.43–45], thereby enabling a wide array of developed tools to be used in the study of the theory.

The second potential infinity comes from the sum over *labels* on each two-complex. There are indications [38.46] that the proper vertex [38.29, 30] introduced in Sect. 38.4.2 may solve this second problem, though, at the moment, these are only indications. Other promising research directions related to this question include [38.47–49]. However, in the following application, we look only at the terms in the sum (38.31) with the lowest power of λ. One can show that the sum over labels for the lowest power of λ is finite, so that, at least for the calculations considered below, this second infinity is not an issue.

38.5.2 The Calculation

To define the two-point correlation function, we first introduce the idea of the *expectation value* of an operator $\hat{O}$ in a given boundary state Ψ, as computed using the path-integral formalism for some region R of space–time. The expectation value is the *average result* one would obtain by measuring the quantity $\hat{O}$ when the system is in the state Ψ. It is denoted $\langle\hat{O}\rangle_\Psi$ and is given by the expression

$$\langle\hat{O}\rangle_\Psi := \frac{\mathcal{A}(\hat{O}\Psi, R)}{\mathcal{A}(\Psi, R)} \,.$$

For illustrative purposes, let us first consider the case of a scalar field theory. In this case, as an aside for those more familiar with standard quantum field theory, when Ψ is an eigenstate $|\varphi(\mathbf{x})\rangle$ of the field operator $\hat{\varphi}(\mathbf{x})$ and $\hat{O}$ is a function $O(\hat{\varphi}(\mathbf{x}))$ of the field operator, the

above expression for the expectation value takes the more familiar form

$$\langle \hat{O} \rangle_\Psi := \frac{\int_{\phi|_{\partial R}=\varphi} O(\phi) e^{iS[\phi]} \mathcal{D}\phi}{\int_{\phi|_{\partial R}=\varphi} e^{iS[\phi]} \mathcal{D}\phi} .$$

Given a canonical quantum state Ψ, and any two points $\mathbf{x}$ and $\mathbf{y}$ on ∂R, the two-point correlation function is defined as

$$D_\Psi(\mathbf{x},\mathbf{y}) := \langle \hat{\varphi}(\mathbf{x})\hat{\varphi}(\mathbf{y})\rangle_\Psi - \langle \hat{\varphi}(\mathbf{x})\rangle_\Psi \langle \hat{\varphi}(\mathbf{y})\rangle_\Psi .$$

In the classical theory, the state of the system is, and therefore uniquely determines, the value of the field $\varphi(\mathbf{x})$ and its conjugate momentum. Thus, classically, given the state of the system, the outcome of a measurement of $\varphi(\mathbf{x})$ is certain, so that the expectation value of $\varphi(\mathbf{x})$ is just $\varphi(\mathbf{x})$, and the expectation value of $\varphi(\mathbf{x})\varphi(\mathbf{y})$ is just $\varphi(\mathbf{x})\varphi(\mathbf{y})$, so that the classical two-point correlation function is just zero. Its deviation from zero can therefore be thought of as a measure of the *nonclassical-ness* of the theory, providing us with an essentially quantum mechanical prediction of the the theory.

The two-point correlation function for loop quantum gravity, as determined by the loop quantum gravity spin-foam model, has been calculated in the works [38.50–54], for both the Euclidean and the Lorentzian versions of the model. The *field operator* in this case (that is, the operator playing the role of $\hat{\varphi}(\mathbf{x})$ above) is the *metric tensor* field operator, which takes a discrete form in the case of loop quantum gravity. The boundary state Ψ that is considered is a linear combination of spin networks based on a fixed graph γ on the boundary of R having the structure indicated in Fig. 38.15.

In order to describe precisely the metric tensor operator and the way it is discretized when acting on the state Ψ, we use again the notion of the complex $\gamma^\star$ dual to γ within the three-dimensional boundary of R. The graph γ within ∂R consists of nodes and links. The dual to each node ν in γ is a three-dimensional region (a 3-cell) $\nu^\star$, and the dual to each link ℓ incident at ν is a two-dimensional surface (a 2-cell) $\ell^\star$ in the boundary of $\nu^\star$. For the specific case of the graph γ given in Fig. 38.15, each 3-cell $\nu^\star$ is a *tetrahedron* and each 2-cell $\ell^\star$ is a *triangle*, so that the dual-cell complex $\gamma^\star$ is a *simplicial* complex in the same sense as Sect. 38.4.1, but one dimension lower, and so provides a *triangulation* of the boundary of R. In fact, $\gamma^\star$ is the boundary of a single 4-simplex (see Sect. 38.4.1); this will be important in a moment.

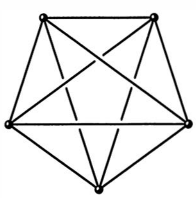


Fig. 38.15 The connectivity of the graph γ used for the boundary state

In terms of the dual-cell complex $\gamma^\star$, the metric tensor operator acting on Ψ is defined as follows. Classically, the metric tensor determines areas and angles. In LQG one has operators corresponding to the areas of the triangles $\ell^\star$ and the interior angles of the tetrahedra $\nu^\star$ in $\gamma^\star$. One can assemble these operators into a single *metric tensor-like matrix* as

$$\hat{h}^{\ell\ell'}(\nu) := \hat{A}_{\ell^\star}\hat{A}_{\ell'^\star}\cos(\hat{\theta}[\nu^\star;\ell^\star,\ell'^\star]) , \qquad (38.32)$$

where $\hat{A}_{\ell^\star}$ is the area operator for the triangle $\ell^\star$, and $\hat{\theta}[\nu^\star;\ell^\star,\ell'^\star]$ is the interior angle between the triangles $\ell^\star$, $\ell'^\star$ within the tetrahedron $\nu^\star$. These areas and angles are precisely the same areas and angles used to interpret LQG spin networks in Sect. 38.2.1–38.2.2 and LQG spin-foam labels in Sect. 38.3.6, except now cast as operators.

We have mentioned that we choose the boundary state Ψ to be based on the graph γ in ∂R. We furthermore choose it to be a *coherent state*, that is, a quantum state which approximates as well as possible a particular *classical state* – one says it is *peaked* on a particular classical state. A classical state in this case consists of an intrinsic geometry of the boundary, described by a matrix $\mathring{h}$ of areas and angles as in (38.32), *together with* a specification $\mathring{\Pi}$ of its conjugate momentum, which, as mentioned in Sect. 38.1.3, describes how ∂R bends in the larger, four-dimensional space R. That is, $\mathring{\Pi}$ describes the *extrinsic* geometry of ∂R. The state Ψ which is used in the calculation is peaked on a particular $\mathring{h}$ and $\mathring{\Pi}$ which are chosen to be as simple as possible, namely they are chosen to describe the intrinsic and extrinsic geometry of the boundary of a regular 4-simplex – that is, a 4-simplex where all of the edges are of equal length.

Now that the nature of the metric tensor field operator and the boundary state has been clarified, we return to the expression for the two-point correlation function. Again, one first defines the notion of the expectation

value of a given operator $\hat{O}$ in the boundary state Ψ

$$\langle \hat{O} \rangle_\Psi := \frac{\mathcal{A}^{\mathrm{LQG}}(\hat{O}\Psi, R)}{\mathcal{A}^{\mathrm{LQG}}(\Psi, R)} ,$$

and the two-point correlation function of the metric tensor operator (38.32) is then

$$G^{\ell_1 \ell_1' \ell_2 \ell_2'}(\nu_1, \nu_2) := \left\langle \hat{h}^{\ell_1 \ell_1'}(\nu_1) \hat{h}^{\ell_2 \ell_2'}(\nu_2) \right\rangle_\Psi - \left\langle \hat{h}^{\ell_1 \ell_1'}(\nu_1) \right\rangle_\Psi \left\langle \hat{h}^{\ell_2 \ell_2'}(\nu_2) \right\rangle_\Psi . \tag{38.33}$$

One can expand this quantity in a power series in the coupling constant λ introduced in Sect. 38.5.1, and what has been so far calculated is the lowest order term in this series, which corresponds to summing over spin foams which include only a single vertex, and which, on the boundary of R, coincide with the graph γ. For the given graph γ in Fig. 38.15, there is only one such spin foam, and its one vertex is dual (in four dimensions) to a single 4-simplex, of which $\gamma^\star$ forms the boundary.

The quantity (38.33) has been calculated in [38.53, 54] (to leading order in λ). It has been found to match, at least in part, the same result one would calculate in a more classical, but incomplete quantum gravity framework – linearized quantum gravity [38.52, 55, 56] – the beginnings of which date back to the work of Rosenfeld, Fierz, and Pauli in the 1930s [38.2]. Linearized gravity is a simplified version of gravity obtained by assuming that space–time geometry is close to flat, so that the metric tensor $g_{\mu\nu}$ is equal to a flat background metric $\eta_{\mu\nu}$ plus some small change $\varepsilon h_{\mu\nu}$, where the components of $h_{\mu\nu}$ are of order one, while ε is much less than one. If one substitutes $g = \eta + \varepsilon h$ into the standard action of gravity $S^{\mathrm{grav}}[\eta + \varepsilon h]$, one can then expand the action in powers of ε. The term with the lowest power of ε, in this case 2, is then the action for linearized gravity. Because the linearized action involves only first and second powers of the basic variable of the theory (usually taken to be h), the theory can be exactly quantized. The two-point correlation function (38.33) calculated using the LQG spin-foam model differs from the two-point correlation function of linearized quantum gravity by addition of a term which goes to zero as the Barbero–Immirzi parameter β goes to zero. This extra term thus yields a new signature of the loop quantum gravity spin-foam dynamics; its significance has yet to be fully understood.

38.6 Discussion

In the spin-foam approach to quantum gravity, one uses what has been learned from canonical loop quantum gravity about quantum *space* to construct a path-integral approach to quantum gravity in which one sums over *quantum space–times*. The resulting framework allows for simpler concrete calculations of the consequences of dynamics than was possible using the canonical methods of loop quantum gravity alone – we have seen this already above in the calculation of the two-point correlation function, and one can also see it in the first steps of the application of the full spin-foam theory to cosmology [38.57–60], a topic which we have not been able to discuss in this chapter.

Beyond these basic developments, the spin-foam approach to quantum gravity has raised other interesting questions and led to further lines of research which are ongoing. These include among others work on how the theory appears on different length scales (so-called *renormalization* of spin foams) [38.61–65], systematic issues in the derivation of spin foams [38.38, 66–69], mathematical tools and equivalent reformulations of spin-foam theory [38.70–72], inclusion of matter [38.73–75], the relation between the dynamics defined by spin-foam sums and the dynamics defined by the Hamiltonian constraint in loop quantum gravity [38.76, 77], and the surprising relation of spin foams to other approaches to quantum gravity, specifically noncommutative geometry [38.78, 79] and group field theory [38.43–45]. These are only a few representative works of the various research directions inspired by spin foams.

If one is to distill a single lesson from the spin-foam program, it is perhaps this: in constructing a path-integral formulation of a quantum theory, it is important to remember the role played by canonical quantization in determining the *potentially discrete nature of the histories one sums over*. A proper path-integral approach to quantum gravity, strictly speaking, should not define transition amplitudes between classical geometries, but rather between *canonical quantum states of quantum gravity*, and one should not sum over classical space–time geometries, but rather *histories of quantum states*.

This is what leads directly to the spin-foam program. In addition to having this firm theoretical basis, the final framework provides a way to combine the advantages of canonical quantum gravity with its predictions of discrete geometry, black-hole entropy, and quantum cosmology, with the manifest unity of space and time made possible by the path-integral approach, a unity of space and time at the heart of both special and general relativity. Lastly, in addition to these theoretical and aesthetic advantages, as we have already touched upon above, the resulting framework allows for simple, concrete calculations involving dynamics by offering an alternative to the task of finding general solutions to the quantum Hamiltonian constraint.

References

38.1 A. Ashtekar, J. Lewandowski: Background independent quantum gravity: A status report, Class. Quantum Gravity **21**, R53 (2004)

38.2 C. Rovelli: *Quantum Gravity* (Cambridge Univ. Press, Cambridge 2004)

38.3 T. Thiemann: *Modern Canonical Quantum General Relativity* (Cambridge Univ. Press, Cambridge 2007)

38.4 J.C. Baez: Spin foam models, Class. Quantum Gravity **15**, 1827–1858 (1998)

38.5 A. Perez (Ed.): The spin-foam approach to quantum gravity, Living Rev. Relativ. 16(3) (2013)

38.6 R. Oeckl: A general boundary formulation for quantum mechanics and quantum gravity, Phys. Lett. B **575**, 318–324 (2003)

38.7 F. Conrady, L. Doplicher, R. Oeckl, C. Rovelli, M. Testa: Minkowski vacuum in background independent quantum gravity, Phys. Rev. D **69**, 064019 (2004)

38.8 H. Leutwyler: Gravitational field: Equivalence of Feynman quantization and canonical quantization, Phys. Rev. **134**, B1155–B1182 (1964)

38.9 J. Hartle, S. Hawking: Wave function of the universe, Phys. Rev. D **28**, 2960–2975 (1983)

38.10 M. Reisenberger, C. Rovelli: Sum over surfaces form of loop quantum gravity, Phys. Rev. D **56**, 3490–3508 (1997)

38.11 P.A.M. Dirac: *Lectures on Quantum Mechanics* (Yeshiva University, New York 1964)

38.12 E. Livine, S. Speziale: A new spinfoam vertex for quantum gravity, Phys. Rev. D **76**, 084028 (2007)

38.13 J.F.G. Barbero: Real Ashtekar variables for Lorentzian signature space-times, Phys. Rev. D **51**, 5507–5510 (1995)

38.14 G. Immirzi: Real and complex connections for canonical gravity, Class. Quantum Gravity **14**, L177–L181 (1995)

38.15 K. Meissner: Black-hole entropy in loop quantum gravity, Class. Quantum Gravity **21**, 5245–5251 (2004)

38.16 I. Agullo, J.F. Barbero, E. Borja, J. Diaz-Polo, E. Villasenor: The combinatorics of the SU(2) black hole entropy in loop quantum gravity, Phys. Rev. D **80**, 084006 (2009)

38.17 H. Ooguri, N. Sasakura: Discrete and continuum approaches to three-dimensional quantum gravity, Modern Phys. Lett. A **6**, 3591–3600 (1991)

38.18 H. Ooguri: Topological lattice models in four-dimensions, Modern Phys. Lett. A **7**, 2799–2810 (1992)

38.19 J. Plebanski: On the separation of Einsteinian substructures, J. Math. Phys. **18**, 2511–2520 (1977)

38.20 J. Barrett, L. Crane: Relativistic spin networks and quantum gravity, J. Math. Phys. **39**, 3296–3302 (1998)

38.21 J. Barrett, L. Crane: A Lorentzian signature model for quantum general relativity, Class. Quantum Gravity **17**, 3101–3118 (2000)

38.22 J. Engle: The Plebanski sectors of the EPRL vertex, Class. Quantum Gravity **28**, 225003 (2011)

38.23 J. Engle: Corrigendum, Class. Quantum Gravity **30**, 049501 (2013)

38.24 J. Barrett, R. Dowdall, W. Fairbairn, H. Gomes, F. Hellmann: Asymptotic analysis of the EPRL four-simplex amplitude, J. Math. Phys. **50**, 112504 (2009)

38.25 J. Barrett, R. Dowdall, W. Fairbairn, F. Hellmann, R. Pereira: Lorentzian spin foam amplitudes: Graphical calculus and asymptotics, Class. Quantum Gravity **27**, 165009 (2010)

38.26 L. Freidel, K. Krasnov: A new spin foam model for 4-D gravity, Class. Quantum Gravity **25**, 125018 (2008)

38.27 J. Engle, E. Livine, R. Pereira, C. Rovelli: LQG vertex with finite Immirzi parameter, Nucl. Phys. B **799**, 136–149 (2008)

38.28 W. Kamiński, M. Kisielowski, J. Lewandowski: Spin-foams for all loop quantum gravity, Class. Quantum Gravity **27**, 095006 (2010)

38.29 J. Engle: A proposed proper EPRL vertex amplitude, Phys. Rev. D **87**, 084048 (2013)

38.30 J. Engle: A spin-foam vertex amplitude with the correct semiclassical limit, Phys. Lett. B **724**, 333–337 (2013)

38.31 E. Buffenoir, M. Henneaux, K. Noui, P. Roche: Hamiltonian analysis of Plebanski theory, Class. Quantum Gravity **21**, 5203–5220 (2004)

38.32 J. Engle, R. Pereira, C. Rovelli: The loop-quantum-gravity vertex-amplitude, Phys. Rev. Lett. **99**, 161301 (2007)

38.33 H. Minkowski: Allgemeine Lehrsätze über die convexen Polyeder, Nachr. Ges. Wiss. Göttingen **1897**(2), 198–219 (1897)

38.34 E. Bianchi, P. Dona, S. Speziale: Polyhedra in loop quantum gravity, Phys. Rev. D **83**, 044035 (2011)
38.35 A. Connelly: Rigidity. In: *Handbook of Convex Geometry*, ed. by P. Gruber, J. Wills (North-Holland, Amsterdam 1993)
38.36 B. Dittrich, S. Speziale: Area-angle variables for general relativity, New J. Phys. **10**, 083006 (2008)
38.37 B. Dittrich, J.P. Ryan: Phase space descriptions for simplicial 4-D geometries, Class. Quantum Gravity **28**, 065006 (2011)
38.38 Y. Ding, C. Rovelli: The volume operator in covariant quantum gravity, Class. Quantum Gravity **27**, 165003 (2010)
38.39 T. Regge: General relativity without coordinates, Nuovo Cim. **19**, 558–571 (1961)
38.40 R. Friedberg, T.D. Lee: Derivation of Regge's action from Einstein's theory of general relativity, Nucl. Phys. B **242**, 392–414 (1984)
38.41 E. Magliaro, C. Perini: Regge gravity from spinfoams, Int. J. Mod. Phys. **22**, 1–21 (2013)
38.42 M. Han, M. Zhang: Asymptotics of spinfoam amplitude on simplicial manifold: Euclidean theory, Class. Quantum Gravity **29**, 165004 (2012)
38.43 R. De Pietri, L. Freidel, K. Krasnov, C. Rovelli: Barrett-Crane model from a Boulatov-Ooguri field theory over a homogeneous space, Nucl. Phys. B **574**, 785–806 (2000)
38.44 D. Oriti: The Group field theory approach to quantum gravity: Some recent results, AIP Conf. Proc., The Planck Scale: Proc. XXV Max Born Symp., Vol. 1196, (Springer, Berlin, Heidelberg 2009) pp. 209–218
38.45 J. Ben Geloun, R. Gurau, V. Rivasseau: EPRL/FK group field theory, Europhys. Lett. **92**, 60008 (2010)
38.46 M. Christodoulou, M. Langvik, A. Riello, C. Roken, C. Rovelli: Divergences, orientation in spinfoams, Class. Quantum Gravity **30**, 055009 (2013)
38.47 C. Perini, C. Rovelli, S. Speziale: Self-energy and vertex radiative corrections in LQG, Phys. Lett. B **682**, 78–84 (2009)
38.48 L. Freidel, D. Louapre: Diffeomorphisms and spin foam models, Nucl. Phys. B **662**, 279–298 (2003)
38.49 L. Freidel, D. Louapre: Ponzano-Regge model revisited I. Gauge fixing, observables and interacting spinning particles, Class. Quantum Gravity **21**, 5685–5726 (2004)
38.50 C. Rovelli: Graviton propagator from background-independent quantum gravity, Phys. Rev. Lett. **97**, 151301 (2006)
38.51 E. Alesci, C. Rovelli: The complete LQG propagator. I. Difficulties with the Barrett-Crane vertex, Phys. Rev. D **76**, 104012 (2007)
38.52 E. Alesci, C. Rovelli: The complete LQG propagator. II. Asymptotic behavior of the vertex, Phys. Rev. D **77**, 044024 (2008)
38.53 E. Bianchi, E. Magliaro, C. Perini: LQG propagator from the new spin foams, Nucl. Phys. B **822**, 245–269 (2009)
38.54 E. Bianchi, Y. Ding: Lorentzian spinfoam propagator, Phys. Rev. D **86**, 104040 (2012)
38.55 M. Veltman: Quantum theory of gravitation, Methods in Field Theory: Les Houches Session XXVIII, ed. by R. Balian, J. Zinn-Justin (World Scientific, Singapore 1981) pp. 265–327
38.56 C. Burgess: Quantum gravity in everyday life: General relativity as an effective field theory, Living Rev. Relativ. **7**, 5 (2004)
38.57 C. Rovelli, F. Vidotto: On the spinfoam expansion in cosmology, Class. Quantum Gravity **27**, 145005 (2010)
38.58 C. Rovelli, F. Vidotto: Stepping out of homogeneity in loop quantum cosmology, Class. Quantum Gravity **25**, 225024 (2008)
38.59 E. Bianchi, C. Rovelli, F. Vidotto: Towards spinfoam cosmology, Phys. Rev. D **82**, 084035 (2010)
38.60 F. Vidotto: Many-nodes/many-links spinfoam: The homogeneous and isotropic case, Class. Quantum Gravity **28**, 245005 (2011)
38.61 C. Rovelli, M. Smerlak: quantum gravity, summing is refining, Class. Quantum Gravity **29**, 055004 (2012)
38.62 B. Dittrich, F.C. Eckert, M. Martin-Benito: Coarse graining methods for spin net and spin foam models, New J. Phys. **14**, 035008 (2012)
38.63 B. Bahr, B. Dittrich, S. Steinhaus: Perfect discretization of reparametrization invariant path integrals, Phys. Rev. D **83**, 105026 (2011)
38.64 V. Rivasseau: Towards Renormalizing Group Field Theory, Proc. Sci. Vol. CNCFG2010 (2010), p. 004
38.65 S. Carrozza, D. Oriti: Bubbles and jackets: new scaling bounds in topological group field theories, J. High Energy Phys. **1206**, 092 (2012)
38.66 J. Engle, M. Han, T. Thiemann: Canonical path integral measures for Holst and Plebanski gravity. I. Reduced phase space derivation, Class. Quantum Gravity **27**, 235024 (2009)
38.67 M. Han: Path-integral for the Master Constraint of Loop Quantum Gravity, Class. Quantum Gravity **27**, 215009 (2010)
38.68 E. Bianchi, D. Regoli, C. Rovelli: Face amplitude of spinfoam quantum gravity, Class. Quantum Gravity **27**, 185009 (2010)
38.69 B. Dittrich, J.P. Ryan: Simplicity in simplicial phase space, Phys. Rev. D **82**, 064026 (2010)
38.70 B. Bahr, F. Hellmann, W. Kaminski, M. Kisielowski, J. Lewandowski: Operator spin foam models, Class. Quantum Gravity **28**, 105003 (2011)
38.71 M. Dupuis, L. Freidel, E.R. Livine, S. Speziale: Holomorphic Lorentzian simplicity constraints, J. Math. Phys. **53**, 032502 (2012)
38.72 M. Dupuis, E.R. Livine: Holomorphic simplicity constraints for 4-D spinfoam models, Class. Quantum Gravity **28**, 215022 (2011)
38.73 E. Bianchi, M. Han, C. Rovelli, W. Wieland, E. Magliaro, C. Perini: Spinfoam fermions, Class. Quantum Gravity **30**, 235023 (2013)

38.74 M. Han, C. Rovelli: Spinfoam fermions: PCT symmetry, Dirac determinant, and correlation functions, Class. Quantum Gravity **30**, 075007 (2013)
38.75 E. Bianchi, T. Krajewski, C. Rovelli, F. Vidotto: Cosmological constant in spinfoam cosmology, Phys. Rev. D **83**, 104015 (2011)
38.76 E. Alesci, C. Rovelli: A regularization of the Hamiltonian constraint compatible with the spinfoam dynamics, Phys. Rev. D **82**, 044007 (2010)
38.77 E. Alesci, T. Thiemann, A. Zipfel: Linking covariant and canonical LQG: New solutions to the Euclidean scalar constraint, Phys. Rev. D **86**, 024017 (2012)
38.78 L. Freidel, E.R. Livine: Ponzano–Regge model revisited III: Feynman diagrams and effective field theory, Class. Quantum Gravity **23**, 2021–2062 (2006)
38.79 L. Freidel, E.R. Livine: Effective 3-D quantum gravity and non-commutative quantum field theory, Phys. Rev. Lett. **96**, 221301 (2006)

39. Loop Quantum Cosmology

Ivan Agullo, Alejandro Corichi

This Chapter provides an up to date, pedagogical review of some of the most relevant advances in loop quantum cosmology. We review the quantization of homogeneous cosmological models, their singularity resolution and the formulation of effective equations that incorporate the main quantum corrections to the dynamics. We also summarize the theory of quantized metric perturbations propagating in those quantum backgrounds. Finally, we describe how this framework can be applied to obtain a self-consistent extension of the inflationary scenario to incorporate quantum aspects of gravity, and to explore possible phenomenological consequences.

39.1 Overview

In this volume there is an introduction to cosmology and cosmic microwave background (CMB) physics by Sourdeep, and on the inflationary paradigm by Wands. They summarize the synergy between theory and observations that has produced spectacular advances in our understanding of the universe in the last decades. The emergence of a *concordance model* is a remarkable success of cosmology and the theory of General Relativity (GR) in which the current paradigm relies. However, the widely accepted Hot Big Bang scenario, regarded as the *standard model of cosmology*, has important limitations, already manifest in its name: the density of matter and the spacetime curvature grow unboundedly in the early universe, blowing up at the big-bang singularity. The big bang is *not* a prediction, but the result of applying the theory *beyond its domain of validity*. When the energy density and curvature approaches the Planck scale, the predictions of General Relativity are unreliable; the *quantum* aspects of the gravitational degrees of freedom are expected to dominate in that regime. This chapter provides a possible quantum gravity extension of the well-established cosmological model from the perspective of loop quantum gravity.

Loop quantum cosmology (LQC) arises from the application of principles of loop quantum gravity (LQG) [39.1, 2] to cosmology. The goal is to quantize the *sector* of General Relativity containing the symmetries of cosmological spacetimes, by following the physical ideas and mathematical tools underlying LQG, presented in detail in Chap. 37. Restricting at-

tention to cosmology presents several advantages. The existence of underlying symmetries largely simplifies technical issues, and allows us to overcome mathematics difficulties that are hard to handle in more generic situations. Yet, the structure is rich enough to contain deep conceptual issues in quantum gravity: What happens with space and time when matter density and curvature reach the Planck scale. Does the big-bang singularity persist? What is the meaning of time in the Planck era? How do classical General Relativity and a smooth spacetime description arise in the low energy regime? What is the scale at which quantum gravity effects become subdominant? Does quantum gravity have anything to contribute to the origin of cosmic structures and to the inflationary scenario? On the other hand, cosmology probably provides the most promising avenue to confront quantum gravity ideas with observations. Cosmology then offers an interesting arena in which quantum gravity can make contact with other theories such as inflation, and probably provides the most promising avenue to confront quantum gravity ideas with observations.

But the restriction to cosmological settings also leads to important limitations. In principle, it is not guaranteed that the result of quantizing a symmetry reduced sector of General Relativity will reproduce the same physics as the restriction of a full quantum gravity theory to symmetric scenarios. Symmetry reduction often entails a drastic simplification, and one may loose important features of the theory by restricting the symmetry prior to quantization. However, it has been extremely useful in several areas of physics, when the complexity of the problem under consideration made it difficult to find solutions without introducing additional inputs. The Oppenheimer–Snyder model of black hole formation and the Dirac quantization of the hydrogen atom are examples that were able to encode the key physical ingredients of the problem, in spite of the severe symmetry reduction. Quantum cosmology may well be another example, if it is constructed choosing carefully the key ingredients from full quantum gravity. It is likely that predictions from quantum cosmology will not agree in every detail with those obtained from full quantum gravity applied to cosmological scenarios, but we expect it to capture the main aspects of the complete theory. As in the previous examples, quantum cosmology can provide valuable information about the correct way to quantize gravity, and be as useful as the hydrogen atom has been for quantum mechanics.

39.1.1 Quantization of Cosmological Spacetimes

General Relativity is a totally constrained theory, in the sense that the full Hamiltonian that generates dynamics is required to vanish. Something similar happens in classical electromagnetism, where *part* of the Hamiltonian, the piece that generates gauge transformations, is a constraint. In General Relativity the constraint turns out to be the *full* Hamiltonian, reflecting the background independence of the theory. Dirac provided the conceptual framework to quantize constrained systems. At the quantum level, physical states have to be annihilated by the operator corresponding to the classical Hamiltonian, $\hat{C}\Psi = 0$, and all the physics has to be extracted from this equation. The quantum state Ψ is the wave function of the physical fields, including the gravitational field itself, and classical quantities such as the metric, energy density and curvature tensor are represented by quantum operators on the physical Hilbert space $\mathcal{H}_{\text{phy}}$ it belongs to. The nontrivial mathematical problem is to make sense and solve the quantum constraint equation, and the underlying cosmological symmetries largely facilitate this task.

The next conceptual issue is to obtain the familiar time evolution that we normally use in physics from this time-less or *frozen* formalism. At the quantum level we do not have a classical metric telling us what are the time-like directions in the manifold, and all what we have is a probability distribution Ψ of different metrics. A useful strategy has been to follow a *relational-time* approach, in which one of the physical variables plays the role of time, and the rest evolve with respect to it. By using a *massless scalar field as this internal time*, it is possible to construct the Hilbert space of physical states satisfying the quantum constraints, and a precise mathematical framework has been developed to study the resulting quantum geometry [39.3]. It has been shown that all the operators representing physical quantities such as the energy density, spacetime curvature, etc., *remain bounded on the physical Hilbert space*, even in the deep Planck regime. This is the mathematical sense in which the singularity is resolved in LQC. The physical picture that emerges from the abstract formalism is the following. When the energy density of the universe is comparable to the Planck energy density, the quantum properties of spacetime geometry become important and dominate. A sort of quantum repulsive degeneracy force appears at such extreme densities, precludes the universe to continue contracting, and forces the quantum spacetime to expand again once the maximum

energy density has been attained, replacing the big-bang singularity by a *quantum bounce*. This maximum energy density is proportional to $\hbar^{-1}$, similar to the finite energy of the ground state of the hydrogen atom that avoids the collapse of the positron and electron as a consequence of the Heisenberg principle. When the energy density and curvature become smaller than approximately 1% of the Planck scale, the quantum effects of gravity become rapidly negligible and classical General Relativity provides an excellent approximation. The resulting quantum dynamics has been analyzed in detail and has provided important insights on the behavior of physics in the Planck regime. The ability of incorporating nonperturbative quantum corrections that are able to completely dominate the evolution in the Planck regime and dilute the big-bang singularity and, at the same time, to disappear in the low energy regime to find agreement with the classical description, is a highly non trivial result of LQC.

Remarkably, some global aspects of the evolution of the quantum geometry can be encoded in simple *effective equation*. Those equations provide a smooth spacetime metric that approximates the full quantum evolution of the quantum spacetime. They have similar form to the equations arising in General Relativity, but include new terms, proportional to $\hbar$, that make the effective trajectory to depart from the classical one around the Planck era. The effective dynamics provides an excellent approximation of the quantum evolution, even at Planckian densities, provided the quantum state is chosen to be highly peaked in a classical trajectory in the low energy regime where General Relativity provides a good approximation.

39.1.2 Inhomogeneous Perturbations in Quantum Cosmology

As emphasized in Chaps. 30 and 32, the theory of inhomogeneous perturbations (of matter and gravitational degrees of freedom) propagating in classical cosmological spacetimes has been a key mathematical tool in modern cosmological research. One of the deepest insights in cosmology is the idea that the cosmic structures (galaxy clusters, superclusters, etc.) that we see today were originated in the very early universe by a process of *amplification of quantum fluctuation by the cosmological expansion*, as explained in the context of cosmic inflation in Chap. 30. In the inflationary scenario, this occurs when the energy density in the universe was close to the grand unification theory (GUT) scale $(10^{16}\,\mathrm{GeV})^4$ around 12 order of magnitude below the Planck energy density. Quantum gravity effects of the background spacetime metric are subdominant at those scales, and the theory of quantized fields propagating in a *classical* background appears to be the appropriate mathematical framework to work out physical predictions. However, earlier in the evolution of the universe, when the curvature and energy density are close to the Planck scale, quantum gravity effects are expected to be important, and they should not be ignored. To have a complete picture of the evolution of cosmic inhomogeneities that encompasses the Planck regime, we need to learn how quantum fields propagate on a *quantum cosmological spacetime* [39.4, 5]. The goal of the second section of this chapter is to review the construction of such a theory.

The detailed description of quantum cosmologies provided by LQC is the suitable arena. The construction of quantum field theory (QFT) on quantum cosmologies follows closely the guiding principle behind LQC: first carry out a truncation of the classical theory adapted to the given physical problem, and then quantize by using LQG techniques. The sector of the classical theory of interest is *extended* in this part to cosmological background *plus first-order inhomogeneous perturbations on it*.

The resulting framework originates from first principles, under the assumption that inhomogeneous perturbations behave as *test fields* on the quantum geometry, and it should provide a bridge between quantum gravity and QFT on curved spacetimes. Therefore, it is suitable to face important conceptual questions such as: What are the concrete approximations under which the familiar QFT in classical spacetimes arises from this more complete description? What are the precise aspects of the quantum geometry that are *seen* by the quantum fields propagating on it? Does the resulting QFT make sense for trans-Planckian modes? These issues will be discussed with some detail in Sect. 39.2. In Sect. 39.3, this framework is applied to the study of gauge invariant cosmic perturbations and phenomenological consequences are worked out.

39.1.3 LQC Extension of the Inflationary Scenario

The inflationary scenario occupies the leading position in accounting for the origin of the cosmic inhomogeneities observed in the CMB and large-scale structure. This success is mainly rooted in the economy of assumptions, the elegant mechanism that originates the *cosmic inhomogeneities from vacuum quan-*

tum fluctuations, namely a subtle interplay between quantum mechanics and classical gravitation, and particularly the nontrivial agreement with observations. Inflation is however an effective theory, and it is expected that a more fundamental theory will complete it. Examples of open questions that the more complete theory should answer are: What is the nature of the scalar inflaton field? Is there a single or several fields, like in multifield models? What is the specific shape of the inflaton potential? These questions originate in particle physics, and unfortunately at these stages LQC does not have much to contribute. There are, in addition, important issues related to gravitation: What is the evolution of the spacetime before inflation? In General Relativity the big-bang singularity is unavoidable in inflationary scenarios [39.6]. Is there a quantum gravity scenario in which the singularity is resolved *and* in which the evolution finds an inflationary phase compatible with observations generically, i. e., *without a fine-tuning of its parameters*? Such a scenario would allow us to extend the inflationary spacetimes all the way back to the Planck era. Moreover, one could then use the quantum theory of cosmological perturbation on quantum spacetimes described in Sect. 39.3, to extend the analysis of cosmic inhomogeneities to include Planck scale physics.

Section 39.4 will review the arguments showing that such an extension is possible in LQC, where one can construct a *conceptual* completion of the inflationary theory from the quantum gravity point of view, in which Planck scale physics can be included in the study of cosmological perturbations. The importance of this extension goes, however, beyond the conceptual domain and may open a window for phenomenological consequences.

To summarize, this chapter will review recent advances in the completion of the quantization program underlying LQG when restricted to the cosmological sector. We shall explore how the singularity of the homogeneous background is avoided, and how the abstract theoretical framework can descend down to make contact with phenomenology. Although many open issues still remain, at the present time there is a solid body of knowledge, based on a rigorous mathematical framework. These combine with analytical and numerical techniques, and provide an avenue from the big-bang singularity resolution to concrete observation of the CMB and galaxy distributions.

Due to space restrictions, there are some topics that we shall not cover in this chapter, such as the path integral formulation and its relation with spin foams [39.7, 8], spin foam cosmology [39.9, 10], and the Gowdy models [39.11–16], nor numerical issues [39.17]. We do not provide either a review of all the existing ideas to study LQC effects on cosmic perturbations. See [39.18–28] for different approaches to that problem. Further information can be found in [39.29–32].

Our convention for the metric signature is $-+++$, we set $c=1$ but keep G and $\hbar$ explicit in our expression, to emphasize gravitational and quantum effects. When numerical values are shown, we use Planck units.

39.2 Quantization of Cosmological Backgrounds

In this section we shall consider the quantum theory of the homogeneous background within the context of LQC. First, we shall discuss what it means for a cosmological model to be quantized, or to use the standard nomenclature, to define a *quantum cosmology*. Just as with the quantization of any mechanical system such as the hydrogen atom, the first step is to cast the model to be quantized in a Hamiltonian language. That is, one has to identify configuration variables q^i and their corresponding momenta p_j, with the property that the Poisson bracket is $\{q^i, p_j\} = \delta^i_j$. The next step in the quantization process is to find a Hilbert space $\mathcal{H}$ and operators $\hat{q}^i$ and $\hat{p}_j$ satisfying $[\hat{q}^i, \hat{p}_j] = i\hbar\delta^i_j$. Then one has to define an operator $\hat{H}$ corresponding to the Hamiltonian (and to other physically relevant observables), in order to define dynamics through the Schrödinger equation: $-i\hbar\partial_t\Psi = \hat{H}\Psi$.

In the case where the classical system under consideration is a *totally constrained system*, instead of a Hamiltonian H defining dynamics, both the classical description and the corresponding quantization are more subtle. Here the dynamical variables are subject to a constraint $C(q,p)=0$. Furthermore, there is no Hamiltonian defining dynamics, and the canonical transformations generated by the constraint C are interpreted as *gauge*. That is, points on the phase space connected by a canonical transformations generated by the constraint are physically equivalent. Thus, the curve on phase space made out of all the physically equivalent points represents a *gauge orbit* and can be

identified with a point on the true, *physical* phase space. Observables will be those functions $f(q,p)$ that are constant along the gauge orbits (i. e., satisfying $\{f, C\} = 0$). Since there is no true dynamics, the system is said to posses a *frozen dynamics*. A natural question is whether one can extract some *dynamics* from the frozen formalism. In some cases, one can use one of the variables (or an appropriately selected function) as an internal time $T(q,p)$, with respect to which the gauge orbit can be described in terms of a relational dynamics (that is, where the *dynamics* is described by correlations between the variable T and the rest of the variables).

Let us now review the quantization process when we have a totally constrained system. The first step is to define a *kinematical Hilbert space* $\mathcal{H}_{\rm kin}$. This space serves an arena for the implementation of the constraint, that is now required to be represented as a self-adjoint operator $\hat{C}$ on $\mathcal{H}_{\rm kin}$. Not all states in the kinematical Hilbert space are regarded as physical. The condition that selects those physical states was put forward by Dirac and has the form

$$\hat{C} \cdot \Psi_{\rm phy} = 0 \,. \tag{39.1}$$

Once one has found the physical states $\Psi_{\rm phy}$ (that might belong to $\mathcal{H}_{\rm kin}$ or not), one needs to specify an inner product $\langle\cdot|\cdot\rangle_{\rm phy}$ in order to construct $\mathcal{H}_{\rm phy}$, the *physical* Hilbert space. Physical observables will be operators $\hat{F}$ that leave the space of physical states invariant. This translates into the condition $[\hat{F}, \hat{C}] = 0$. In some cases, when there is an internal time variable T, one can recast the Dirac condition (39.1) as an *evolution* equation where T plays the role of time, as in the Schrödinger equation.

One interesting feature of the simplest cosmological models is that they are totally constrained systems, so the general framework we have outlined is applicable. Even more, one can complete the quantization program and obtain a complete physical description where a massless scalar field ϕ plays the role of internal relational time. One can then pose physical questions pertaining to observables of cosmological interest, such as the Hubble parameter and curvature scalars. Interestingly, for the simplest models, one can indeed find *two* different, inequivalent, quantizations. The first one corresponds to the so-called Wheeler–De Witt (WDW) quantization that was put forward by De Witt and Misner in the 60s. The second quantizations corresponds precisely to the one we shall here describe in detail, known as LQC. As we shall describe in more detail later, the basic difference between these two programs corresponds to the choice of kinematical Hilbert space $\mathcal{H}_{\rm kin}$. The choice made by De Witt and others was, in a sense, the most natural one, resembling the Schrödinger quantum mechanics that has been very useful to describe many physical systems. On the other hand, the choice one makes in LQC is somewhat exotic from the perspective of standard quantum mechanics, but is selected when the underlying symmetries pertinent to the gravitational field are seriously taken into account.

The second and physically most important difference between these two representations is that their predictions regarding the fate of the classical singularity are radically different. While the WDW theory predicts that the singularity remains, as defined by the behavior of the expectation values of physically relevant operators such as energy density, in the case of LQC the singularity is generically avoided. Instead of a big bang (or big crunch) one has a bounce connecting a contracting branch with an expanding one; the energy density and curvature scalars are bounded from above, so that physics is well-defined throughout the intrinsic dynamical evolution of the quantum state describing the universe.

Let us now briefly describe the structure of the remainder of this section. In the first part, we study in detail the $k=0$ Friedmann–Lemaître–Robertson–Walker (FLRW) model with vanishing cosmological constant and discuss some of its main features. In the second part we discuss other models. The first one we consider is the closed $k=1$ model also without a cosmological constant. Next, we briefly discuss $k=0$ FLRW models with a cosmological constant and some anisotropic models. In the third part, we introduce the so called effective equations. We give a brief introduction to the subject and discuss in detail the case of the $k=0$ FLRW model. Next we consider the $k=1$ case, followed by a discussion of anisotropic effective spacetimes, including the Bianchi I, II, and IX models.

39.2.1 $k = 0$ FLRW, Singularity Resolution

The simplest model that one can consider is a $k=0$ homogeneous and isotropic FLRW cosmological model foliated by 3-manifolds Σ that are topologically $\mathbb{R}^3$. In order to find a Hamiltonian description for the model, we have to start with an action principle. Due to homogeneity, the action is not well defined unless one introduces and fixes a fiducial cell $\mathcal{V}$. This will play the role of a comoving volume. We can introduce a flat fiducial metric $\mathring{q}_{ab}$ on $\mathbb{R}^3$ with respect to which the co-

ordinate volume of $\mathcal{V}$ is $\mathring{V} = \int_{\mathcal{V}} \sqrt{\mathring{q}}\,\mathrm{d}^3x$ without loss of generality, we shall set $\mathring{V} = 1$. The flat FLRW spacetime is described by the metric

$$\mathrm{d}s^2 = -N^2\,\mathrm{d}t^2 + a(t)^2\,\mathrm{d}\mathbf{x}^2 \,, \qquad (39.2)$$

where N is the lapse function, $\mathring{q} \leftrightarrow \mathrm{d}\mathbf{x}^2$ is the flat fiducial metric, and a is the *scale factor* of the universe. Now, the action principle is

$$\begin{aligned} S &= \frac{1}{16\pi G} \int \mathrm{d}t \int_{\mathcal{V}} \mathrm{d}^3x \sqrt{|g|} R \\ &= \frac{1}{16\pi G} \int \mathrm{d}t N a^3 R \,, \end{aligned}$$

with R the scalar curvature of the spacetime. The gravitational part of the phase space consists of a and its conjugate momenta that is found to be

$$P_a = -\frac{3}{4\pi G N} a\dot{a} \,.$$

In this simplest model, the matter we shall consider is a homogeneous massless scalar field ϕ. The action for such a field is

$$S_{\mathrm{matt}} = \frac{1}{2} \int \mathrm{d}t \frac{a^3 \dot{\phi}^2}{N} \,.$$

From this, the momenta $p_{(\phi)}$ associated to the scalar field is $p_{(\phi)} = (\dot{\phi} a^3)/N$, and the Hamiltonian constraint that defines the *dynamics* is then

$$C_{\mathrm{tot}} = \frac{2\pi G}{3} \frac{P_a^2}{a} - \frac{1}{2} \frac{p_{(\phi)}^2}{a^3} \approx 0 \,. \qquad (39.3)$$

To summarize, the phase space is four dimensional with coordinates $(a, P_a; \phi, p_{(\phi)})$, satisfying $\{a, P_a\} = 1$ and $\{\phi, p_{(\phi)}\} = 1$. In the standard WDW approach, the next step is to consider the kinematical Hilbert space to consist of *wave functions* $\Psi_{\mathrm{wdw}} = \Psi(a, \phi)$ of the *configuration* variables (a, ϕ). In this case, the operators are represented in the usual fashion, as: $\hat{a} \cdot \Psi(a, \phi) = a\Psi(a, \phi)$ and $\hat{P}_a = -\mathrm{i}\hbar \partial_a \Psi(a, \phi)$, and similarly for the other variables. Then, one promotes the constraint (39.3) to an operator, and finds solutions to the Dirac condition (39.1). This has been described in detail in [39.33, 34].

In order to define the corresponding phase space in LQC, we need to follow some more steps. The first one is that one needs to introduce a new set of variables for the gravitational degrees of freedom. As explained in Sahlmann's contribution to this volume, LQG, and consequently LQC is based in a connection A and its corresponding momenta E, a generalization of the magnetic potential and electric field of electromagnetism.

Due to the underlying symmetries of the spacetimes we are considering, these variables can be written as [39.3]

$$A_a^i = \tilde{c}\,\mathring{\omega}_a^i \,; \quad E_i^a = \tilde{p}\sqrt{\mathring{q}}\,e_i^a \,, \qquad (39.4)$$

where e_i^a is a fiducial triad and $\mathring{\omega}_a^i$ is the cotriad compatible with $\mathring{q}_{ab}$. Now, the dynamical variables in the isotropic cosmological regime are p and c. The relationship between the *triad* p and the scale factor is, $|p| = a^2$. The connection component gets related to the rate of change of scale factor as $c = \gamma \dot{a}/N$, holding only for the physical solutions of general relativity. For convenience during the loop quantization process, let us introduce the variables $V = a^3$, the volume, and its conjugate variable $b := c/|p|^{1/2}$. The gravitational part of the phase space is characterized by the conjugate variables V and b satisfying

$$\{b, V\} = 4\pi G \gamma \,, \qquad (39.5)$$

and the complete phase space has coordinates $(b, V; \phi, p_{(\phi)})$. A further simplification is to choose $N = a^3 = V$ from the very beginning. If we rewrite the line element with this choice we have $\mathrm{d}s^2 = -a^6\,\mathrm{d}\tau + a^2\,\mathrm{d}\mathbf{x}^2$, for which the classical constraint now reads

$$\tilde{C} = p_{(\phi)}^2 - \frac{3}{4\pi G \gamma^2} V^2 b^2 = 0 \,. \qquad (39.6)$$

It is worthwhile to note that the dynamics thus fund in the Hamiltonian language is completely equivalent to the standard description based in Einstein's equations.

Let us now consider the issue of quantization. As previously discussed, the choice of kinematical Hilbert space in LQC is different from the WDW case. That is, we do not expect to represent $\hat{b}$ and $\hat{V}$ as multiplication and derivation, for example. The idea instead is to construct a quantum theory that is closest to the quantization used in loop quantum gravity, as discussed in Sahlmann's contribution. This means in particular a different choice of kinematical Hilbert space. Recently this *polymeric* quantization for cosmological models has been shown to be unique when invariance under diffeomorphisms is imposed [39.35] (in complete analogy with the corresponding results in full LQG [39.36, 37]). The new strategy is the following. Instead of re-writing

the Hamiltonian constraint (39.3) in terms of the (b, V) variables, one starts the full expression of the Hamiltonian constraint, in terms of variables A and E. Then, one uses the simplification given by (39.4). As it turns out, the choice of the polymeric Hilbert space as the kinematical arena for the implementation of the constraint – following the LQG route to quantization – has the important feature that it does *not* admit the $\hat{b}$ operator. That is, only exponential functions of the variable b such as

$$h^{(\lambda)} = \exp\left(\mathrm{i}\lambda\frac{b}{2}\right) \tag{39.7}$$

become well defined. These functions generate an algebra of so-called almost periodic functions, for arbitrary λ.

The basic assumption behind loop quantization is that the corresponding functions become a basis for the quantum theory. The resulting kinematical Hilbert space is then $L^2(\mathbb{R}_{\mathrm{Bohr}}, \mathrm{d}\mu_{\mathrm{Bohr}})$, a space of square integrable functions on the *Bohr compactification* of the real line. It is straightforward to understand the nature of this space. For instance, the eigenstates of $\hat{V}$, labelled by $|\nu\rangle$, satisfy $\langle\nu_1|\nu_2\rangle = \delta_{\nu_1,\nu_2}$. This is to be contrasted with the usual Schrödinger representation where, instead of the Kronecker delta, one has the Dirac delta. This representation of quantum states and operators is referred to as the *polymer representation* because in full LQG the fundamental excitations of the gravitational field are one dimensional and polymer like.

In particular, these eigenstates are *normalized* and constitute a basis for the kinematical Hilbert space $\mathcal{H}_{\mathrm{poly}}$. This constitutes the main difference from the standard Schrödinger representation where the eigenstates of momentum $\hat{p}|\nu\rangle = \nu|\nu\rangle$ are *not* normalized and satisfy $\langle\nu|\mu\rangle = \delta(\mu, \nu)$. Note also that this plane waves states are *not* a basis for the $L^2(\mathbb{R}, \mathrm{d}x)$ Hilbert space.

There exists an important result in mathematical physics stating that for a finite-dimensional phase space, the Schrödinger Hilbert space is the only choice of representation of the canonical commutation relations, satisfying some regularity conditions. This result goes under the name of the Stone–Von Neumann uniqueness theorem [39.38]. Thus, one could have imagined that, since the system has a finite number of degrees of freedom, both the WDW and the LQC representations should be equivalent. However, that expectation is not realized. The polymeric representation used in LQC and the standard one are unitarily inequivalent. This is due to a crucial property of the LQC operators, implying that the polymer quantum mechanics does not posses some of the regularity conditions that go into the hypothesis of the Stone–Von Neumann theorem. To explore those properties further, let us consider the action of the two fundamental operators on the eigenstates $|\nu\rangle$,

$$\hat{V}|\nu\rangle = 2\pi\gamma\ell_{\mathrm{Pl}}^2\nu|\nu\rangle\,; \quad \widehat{\exp\left(\mathrm{i}\lambda\frac{b}{2}\right)}|\nu\rangle = |\nu+\lambda\rangle\,. \tag{39.8}$$

Note that the *displacement* operator $\widehat{\exp(\mathrm{i}\lambda b/2)}$ is not continuous when $\lambda \to 0$, since the states $|\nu\rangle$ and $|\nu+\lambda\rangle$ are always orthogonal to each other, for all nonzero values of λ. Also note that a basis of the polymer Hilbert space is uncountable as the label ν for the eigenstates can take any value in the real line.

Let us now find what the form of the quantum constraint operator is. The idea is to consider wave functions of the type $\tilde{\Psi}(\nu, \phi) \in \mathcal{H}_{\mathrm{kin}}$. The quantum constraint operator on wave functions $\tilde{\Psi}(\nu, \phi)$ of ν and ϕ is then

$$\partial_\phi^2\tilde{\Psi}(\nu, \phi) =: \Theta_{(\nu)}\tilde{\Psi}(\nu, \phi)\,. \tag{39.9}$$

The geometrical part, $\Theta_{(\nu)}$, of the constraint is a difference operator in steps of 4λ, that takes the form

$$\begin{aligned}\Theta_{(\nu)} &:= C^+(\nu)\Psi(\nu+4\lambda) + C^0\Psi(\nu)\\ &\quad + C^-\Psi(\nu-4\lambda)\Psi(\nu)\,,\end{aligned} \tag{39.10}$$

where $C^\pm$ and C^0 are functions of $|\nu|$ [39.34, 39]. Note that the equivalent of the WDW equation is now a *difference* equation in the geometrical variable, instead of a differential equation.

Then, physical states correspond to solutions to the quantum constraint (39.9), but they should also belong to the positive frequency part of the Hamiltonian constraint, and satisfy the *Schrödinger equation*

$$-\mathrm{i}\hbar\partial_\phi\Psi(\nu, \phi) = \sqrt{\Theta}\Psi(\nu, \phi) \equiv H_0\Psi(\nu, \phi)\,. \tag{39.11}$$

Furthermore, they should be symmetric under $\nu \to -\nu$ and have finite norms under the inner product

$$\langle\Psi_1|\Psi_2\rangle = \sum_\nu \overline{\Psi}_1(\nu, \phi_0)|\nu|^{-1}\Psi_2(\nu, \phi_0)\,, \tag{39.12}$$

where the constant ϕ_0 is arbitrary. As discussed above, these physical states can be interpreted as being solu-

tions to *evolution equations* with respect to the internal time ϕ.

The next step is to define relational observables that will have a clear interpretation in terms of ϕ. For instance, one can define the operator $\hat{V}_{\phi_0}$ as the operator corresponding to *the volume V when the scalar field takes the value* ϕ_0. One can indeed define such Heisenberg operators by the standard prescription

$$\hat{V}|_{\phi_0} \cdot \Psi_{\text{phy}}(\nu,\phi) := e^{iH_0(\phi-\phi_0)}\hat{V}e^{-iH_0(\phi-\phi_0)} \times \Psi_{\text{phy}}(\nu,\phi)\,, \tag{39.13}$$

where $\hat{V}$ is the standard Schrödinger operator (acting by multiplication in this case). In this manner one can define operators corresponding to matter energy density $\hat{\rho}_{\phi_0}$ and curvature scalars, all with a clear interpretation as being defined at *time* ϕ_0.

As it turns out, one can perform a Fourier transform into the conjugate variable to ν, and the resulting quantum constraint, a differential equation, can be solved exactly [39.34]. This allows one to have closed expressions for the expectation values of the Heisenberg operators. For instance, the expectation value for the volume operator takes the form

$$\langle\hat{V}\rangle_\phi = V_+ e^{\alpha\phi} + V_- e^{-\alpha\phi}\,, \tag{39.14}$$

with, $V_\pm$ constants that depend on the details of the initial (normalized) wave function, and $\alpha = \sqrt{12\pi G}$. From (39.14), it follows that the expectation value of the volume $\hat{V}|_\phi$ is large at both very early and late times and has a nonzero global minimum

$$V_{\min} = 2(V_+V_-)^{1/2}\,.$$

The *bounce* occurs at time

$$\phi_b^V = (2\alpha)^{-1}\ln\left(\frac{V_-}{V_+}\right)\,.$$

Around $\phi = \phi_b^V$, the expectation value of the volume $\langle\hat{V}\rangle_\phi$ is symmetric. Thus we see that *all* states undergo a *big bounce* that replaces the big bang (in which the volume goes to zero as $\phi \to \pm\infty$). Note: In the case of the WDW quantization, the expected volume reaches zero as $\phi \to \pm\infty$, so in this sense one still reaches the singularity.

Another important observable to consider is the energy density $\hat{\rho}|_\phi$. Interestingly, this quantity possesses an absolute upper bound given by

$$\langle\hat{\rho}\rangle_\phi \le \rho_{\max} \quad \text{with} \quad \rho_{\max} := \frac{3}{8\pi\gamma^2 G}\frac{1}{\lambda^2}\,. \tag{39.15}$$

It is interesting to note that this quantity depends inversely with the *loop quantum geometry scale* λ. Thus, in the limit $\lambda \to 0$, where we expect to recover the WDW theory, the density becomes unbounded. That is precisely what is found in the complete quantization of the WDW theory [39.34].

Using the standard choice for λ in LQC, namely $\lambda^2 = 4\pi\sqrt{3}\gamma\ell_{\text{Pl}}^2$, we obtain

$$\rho_{\max} = \frac{\sqrt{3}}{32\pi^2\gamma^3 G^2\hbar} \approx 0.41\rho_{\text{Pl}}$$

(using the standard choice for γ in LQG).

Let us now summarize the main features of the complete quantization of this simple cosmological model.

1. The bounce is not restricted to semiclassical states but occurs for states in a dense subspace of the physical Hilbert space.
2. There exists a supremum of the expectation value for the energy density. This maximum allowed density is $\rho_{\max} = \sqrt{3}/(32\pi^2\gamma^3 G^2\hbar)$. We note that existence of an absolute maximum of the energy density in this cosmological model implies a nonsingular evolution, in terms of physical quantities. The singularity is therefore, resolved.
3. When curvatures become much smaller than the Planck curvature (or for $\rho \ll \rho_{\max}$), the expectation values of the Dirac observables agree with the values obtained from classical GR.
4. For states which are semiclassical at late times, i.e., those which lead to a large classical universe, the backward evolution leads to a quantum bounce in which the energy density of the field becomes arbitrarily close to $\rho_{\max} \approx 0.41\rho_{\text{Pl}}$.
5. States that evolve to be semiclassical at late times, as determined by the dispersion in canonically conjugate observables, have to evolve from states that also had semiclassical properties before the bounce (even when there might be asymmetry in their relative fluctuations without affecting semiclassicality) [39.40–43]. Semiclassicality is preserved to an amazing degree across the bounce.

This concludes our discussion of the quantization of the homogeneous background in the case where the matter content is a massless scalar field. This is the simplest isotropic model and is completely solvable. The

question now is how to generalize these results for other isotropic and anisotropic models. That will be subject of the next section.

39.2.2 Other Cosmologies

$k = 1$ FLRW

There are several generalizations one might consider away from the $k = 0$, $\Lambda = 0$, FLRW cosmology. The simplest case is to consider the $k = 1$ FLRW cosmological model [39.44–46]. Even when it is not phenomenologically favored, it is important since it represents a spatially closed model that in the classical theory has both an expanding and a contracting phase continuously joined by a *recollapse* point where $H = \dot{a}/a = 0$. Therefore, it is an important test if one can recover the classical recollapse from the quantum theory.

The spacetimes under consideration are of the form $M = \Sigma \times \mathbb{R}$, where Σ is a topological three-sphere $\mathbb{S}^3$. It is standard to endow Σ with a fiducial basis of one-forms ${}^o\omega^i_a$ and vectors ${}^oe^a_i$. The fiducial metric on Σ is then ${}^oq_{ab} := {}^o\omega^i_a\, {}^o\omega^j_b k_{ij}$, with k_{ij} the Killing–Cartan metric on su(2). Here, the fiducial metric ${}^oq_{ab}$ is the metric of a three-sphere of radius a_0. The volume of Σ with respect to ${}^oq_{ab}$ will be denoted by $V_0 = 2\pi^2 a_0^3$. We also define the quantity $\ell_0 := V_0^{1/3}$. It can be written as $\ell_0 =: \vartheta a_0$, where the quantity $\vartheta := (2\pi^2)^{1/3}$ will appear in many expressions.

The isotropic and homogeneous connections and triads can be written in terms of the fiducial quantities as follows:

$$A^i_a = \frac{c}{\ell_0}\,{}^o\omega^i_a\,; \quad E^a_i = \frac{p}{\ell_0^2}\sqrt{{}^oq}\,{}^oe^a_i\,. \tag{39.16}$$

Here, c is dimension-less and p has dimensions of length. The metric and extrinsic curvature can be recovered from the pair (c, p) as follows:

$$q_{ab} = \frac{|p|}{\ell_0^2}\,{}^oq_{ab}\,,$$

and

$$\gamma K_{ab} = \left(c - \frac{\ell_0}{2}\right)\frac{|p|}{\ell_0^2}\,{}^oq_{ab}\,.$$

Note that the total volume V of the hypersurface Σ is given by $V = |p|^{3/2}$. The only relevant constraint is the Hamiltonian constraint that has the form

$$C_{\text{grav}} = -\frac{3}{8\pi G\gamma^2}\sqrt{|p|}\left[(c - \vartheta)^2 + \gamma^2\vartheta^2\right]. \tag{39.17}$$

It is convenient to also use the variables [39.34]: $b := c/|p|^{1/2}$ and $V = p^{3/2}$. The quantity V is just the volume of Σ and b is its canonically conjugate, $\{b, V\} = 4\pi G\gamma$. We can then compute the evolution equations of V and b in order to find interesting geometrical scalars. Then

$$\dot{V} = \{V, C_{\text{grav}}\} = \frac{3}{\gamma}\left(bV - \vartheta V^{2/3}\right), \tag{39.18}$$

from which we can find the standard Friedman equation using the constraint equation $C = C_{\text{grav}} + C_{\text{matt}} \approx 0$ and $C_{\text{matt}} = V\rho$, we have

$$H^2 := \left(\frac{\dot{V}}{3V}\right)^2 = \frac{8\pi G}{3}\rho - \frac{\vartheta^2}{V^{2/3}}\,.$$

The basic strategy of loop quantization, just as in the $k = 0$ case, is that the effects of quantum geometry are manifested by means of holonomies around closed loops to carry information about field strength of the connection. In order to define the quantum theory, taking again $N = a^3$, one can work in the ν representation and define operators associated to curvature and spin connection to arrive at a difference operator $\Theta_{(k=1)}$ of the form

$$\begin{aligned}
&\partial_\phi^2 \Psi(\nu, \phi) \\
&= -\bullet_{(k=0)}\Psi(\nu, \phi) \\
&\quad + \frac{3\pi G}{\lambda^2}\nu\left[\sin^2\left(\frac{\lambda\vartheta}{\tilde{K}\nu^{1/3}}\right) + (1 + \gamma^2)\left(\frac{\lambda\vartheta}{\tilde{K}}\right)\right] \\
&\quad \times \Psi(\nu, \phi)\,,
\end{aligned} \tag{39.19}$$

with $\tilde{K} = 2\pi\gamma\ell_{\text{Pl}}$. Numerical solutions of this equation were studied in detail in [39.44] for sharply peaked states, and were shown to posses not only a bounce very close to the critical density $\rho_{\max}$, but also a recollapse at a density and volume very close to the classical value. Thus, this model provides a very striking example of a quantum gravitational system that possesses satisfactory ultraviolet (UV) and infrared (IR) behavior. The relative dispersion of $\hat{V}|_\phi$ does increase but the increase is very small: For a universe that undergoes a classical recollapse at ≈ 1 Mpc, a state that nearly saturates the uncertainty bound initially, with uncertainties in $\hat{p}_\phi$ and $\hat{V}|_\phi$ spread equally, the relative dispersion in $\hat{V}|_\phi$ is still $\approx 10^{-6}$ after some 10^{50} cycles [39.44]. The expectation values of volume have a quantum bounce which occurs at $\rho = \rho_{\max}$ up to the correction terms of the order

of $\ell_{\text{Pl}}^2/V_{\text{bounce}}^{2/3}$. For universes that grow to macroscopic sizes, the correction is totally negligible. For example, for a universe which grows to a maximum volume of 1 Gpc^3, the volume at the bounce is approximately $10^{117}\ell_{\text{Pl}}^3$. On the other hand, the numerical simulations show that one indeed recovers the recollapse with very large precision for semiclassical states that reach large volumes [39.44]. An important lesson that this model teaches us is that energy density and curvature are the relevant quantities to define what the Planck scale is, and not the size of the universe at the bounce (that, as we have seen, can be very large in Planck units). One should also note that, while semiclassical states alternate between the Planck scale (UV) and the low density, large volume GR regime (IR) states that are *truly quantum* – or far from semiclassical – might have a bounce at a density much lower than ρ_{max}, and not grow to large volumes before recollapse.

There exists another quantization in which the curvature is not obtained by means of closed holonomies, but rather by approximating the *connection* by open holonomies, as is done in anisotropic models with nontrivial curvature [39.46]. The structure of the constraints is different but its quantum solutions have not been explored numerically yet.

Let us comment on the quantization of the $k = -1$ case. Some early attempts to find such a quantization were put forward in [39.47, 48], but those efforts still suffer from some drawbacks, such as the absence of essential self-adjointness. A quantization based in open holonomies as in [39.46] is still to be constructed.

FLRW with $\Lambda \neq 0$

The results found for a zero cosmological constant can be generalized to the case of a nonzero cosmological constant. For a mass-less scalar field and both signs of the constant, we have singularity resolution, in the sense that the big bag/crunch is replaced by a bounce, just as in the $\Lambda = 0$ case. For simplicity we shall consider the $\Lambda < 0$, $k = 0$ case, but the results can be generalized to $k = 1$ as well. The Hamiltonian constraint, for $N = 1$, takes the form

$$C = \frac{p_{(\phi)}^2}{2V} - \frac{3}{8\pi G\gamma^2} b^2 V + \frac{\Lambda}{16\pi G} V \approx 0 \,. \quad (39.20)$$

One can solve the equations of motion and express the dynamics in terms of the scalar field ϕ as

$$V(\phi) = \frac{\alpha p_{(\phi)}}{\sqrt{3|\Lambda|}} \frac{1}{\cosh[\alpha(\phi - \phi_0)]} \,. \quad (39.21)$$

With this, there is a big-bang singularity in the past $\phi \to -\infty$ and a big crunch in the future, when $\phi \to \infty$. There is a point of recollapse, when the volume reaches its maximum value $V_{\text{max}} = (\alpha p_{(\phi)})/(\sqrt{3|\Lambda|})$, at $\phi = \phi_0$, with some resemblance to the $k = 1$ case. The quantum constraint now takes the form

$$\partial_\phi^2 \Psi(\nu, \phi) = -\Theta \Psi(\nu, \phi) - \frac{\pi G\gamma^2 |\Lambda|}{2} \nu^2 \Psi(\nu, \phi) \,, \quad (39.22)$$

with Θ the operator corresponding to the $k = 0$, $\Lambda = 0$ case. The operator can be consistently defined, and numerically solved [39.49] to give a picture very similar to the $k = 1$ case with vanishing cosmological constant. The big bang/crunch is replaced by a bounce, in such a way that a sharply peaked state goes through a series of bounces and recollapses in an almost periodic fashion.

Let us now consider the $\Lambda > 0$ case. The solution to the classical equations is slightly different from the negative case and takes the form [39.50, 51]

$$V(\phi) = \frac{\alpha p_{(\phi)}}{\sqrt{3|\Lambda|}} \frac{1}{\sinh[\alpha(\phi - \phi_0)]} \,. \quad (39.23)$$

This is qualitatively very different from the previous case. Now, an expanding solution with a big-bang singularity at the past, $\phi \to -\infty$, reaches an infinite volume for a *finite* value of ϕ, namely when $\phi = \phi_0$. Similarly, there are contracting solutions that *start*, for $\phi = \phi_0$, with an infinite volume and end in a big crunch singularity when $\phi \to \infty$. At the point $\phi = \phi_0$, the proper time diverges and the matter density vanishes. One can see that one can actually continue the classical evolution past this *singular* point [39.50]. In the quantum theory, this new behavior manifests itself in the fact that the operator Θ_Λ fails to be essentially self-adjoint, and one has the freedom of choosing different self-adjoint extensions. Interestingly enough, for all of them, the evolution of semiclassical states is almost indistinguishable. Evolution is well defined past the point $\phi = \phi_0$ and the universe recollapses. As in all previous cases, the big bang/crunch singularity is replaced by a bounce.

Anisotropic Cosmologies

Isotropic LQC, as we have seen, enjoys a very robust formulation; one has complete mathematical control over the quantum theory, one can make physical predictions using analytical or numerical tools and

can therefore draw conclusions about the behavior of a background isotropic quantum geometry. The same is not true for anisotropic solutions. While the quantum constraints have been formulated in several cases, one does not have full mathematical control regarding their time evolution, and one has not been able to solve, even numerically, their dynamical evolutions. In this part we shall summarize the formulation of the quantum models as we currently understand them.

Let us consider the spacetime of the form $M = \Sigma \times \mathbb{R}$ where Σ is a spatial 3-manifold which can be identified by the symmetry group of the chosen model and is endowed with a fiducial metric ${}^o q_{ab}$ and associated fixed fiducial basis of 1-forms ${}^o\omega^i_a$ and vectors ${}^o e^a_i$. If Σ is noncompact then we fix a fiducial cell $\mathcal{V}$ adapted to the fiducial triads with finite fiducial volume. We also define L_i which is the length of the ith side of the cell along ${}^o e_i$ and $\mathring{V} = L_1 L_2 L_3$. We choose for compact Σ, $L_i = \mathring{V}^{1/3}$ with $i = 1, 2, 3$.

Since all of the models in which we are interested are homogeneous and, if we restrict ourselves to diagonal metrics, one can fix the gauge in such a way that A^i_a has three independent components, c^i, and E^a_i has three independent components, p_i

$$A^i_a = \frac{c^i}{L_i}{}^o\omega^i_a \quad \text{and} \quad E^a_i = \frac{p_i L_i}{\mathring{V}}\sqrt{{}^o q}\,{}^o e^a_i\,, \tag{39.24}$$

where p_i, in terms of the scale factors a_i, are given by $|p_i| = L_j L_k a_j a_k$ $(i \neq j \neq k)$. Using (c^i, p_i) for anisotropic models, the Poisson brackets can be expressed as $\{c^i, p_j\} = 8\pi G\gamma\delta^i_j$. With this choice of variables and gauge fixing, the Gauss and diffeomorphism constraints are automatically satisfied and, again, only the Hamiltonian constraint remains.

In the Bianchi I model, the quantum constraint operator can be constructed by a natural generalization of the strategy used in the isotropic case [39.52–54]. In the Bianchi II and IX models, however, an extension of the strategy is needed [39.55, 56]. In the $k = 0$ and Bianchi I models one can also use the new strategy and it yields the same quantum Hamiltonian constraint. However, for the $k = 1$ isotropic models the two strategies yield different quantum constraints – reflecting a quantization ambiguity – and the quantum constraint obtained by the new method is the limit of the Bianchi IX quantum constraint in the isotropic limit [39.46, 56, 57].

By using these results and choosing some factor ordering, we can construct the total constraint operator. Note that different choices of factor ordering will yield different operators, but the main results will remain almost the same. By solving the constraint equation $\hat{C}_H \cdot \Psi = 0$, we can obtain the physical states and the physical Hilbert space $\mathcal{H}_{\text{phys}}$. As a final step, one would need to identify the physical observables, that in our case would correspond to relational observables as functions of the internal time ϕ. These steps have proven to be exceptionally difficult and so far these difficulties have prevented from solving the resulting difference equations numerically, even for the simplest case of Bianchi I.

39.2.3 Effective Equations

When analyzing the numerical solutions of the $k = 0$, $\Lambda = 0$ FLRW model, the authors of [39.33] noticed that sharply peaked states followed trajectories in the (V, ϕ) plane that have a bounce, and therefore do not satisfy the classical Einstein equations. Furthermore, they realized that the expectation value of $\hat{V}|_\phi$ *does indeed* follow a trajectory that satisfies (to a very good approximation) some so-called *effective equations*. As it turns out, these effective equations can be derived from an effective Hamiltonian constraint C_{eff}. The question that arises then is how to derive, from the quantum theory defined by a quantum constraint $\hat{C}$, the effective Hamiltonian. A second question pertains to the domain of validity of these effective equations. That is, for which states and in which regimes are these equations a good approximation to the exact quantum dynamics? As we shall see in this part, for the models that are well understood, effective equations describe very accurately the dynamics for appropriately defined semiclassical states.

In the case of models for which we do not posses the full quantum dynamics, one can expect that the effective theory to describe very well the quantum theory for semiclassical states far from the *deep quantum regime* (where it is expected to fail). Thus, in the anisotropic Bianchi I, II, and IX models, the effective description that we shall here consider provide a description in which the singularity is also replaced by a bounce.

Let us begin by briefly describing how one obtains this effective descriptions from the quantum theory. The idea is to employ the geometric formulation of quantum mechanics [39.58], which provides an appropriate formalism from which one can find the effective Hamiltonian constraint C_{eff} by computing the expectation value $\langle\hat{C}\rangle_\psi$ of the quantum Hamiltonian constraint on an appropriately defined semiclassical state ψ. From that expression one can find the effective equations of motion by replacing C_{eff} in Hamilton's equations: $\dot{q} = \{q, C_{\text{eff}}\}$ and $\dot{p} = \{p, C_{\text{eff}}\}$.

Let us now consider some important cases in homogeneous LQC.

$k = 0$ FLRW cosmology

Using the geometric methods of quantum mechanics, one can write an effective Hamiltonian which provides an excellent approximation to the behavior of expectation values of Dirac observables in the numerical simulations [39.59]. The effective Hamiltonian will, in principle, also have contributions from terms depending on the properties of the state such as its spread. The effect of these terms turns out to be negligible as displayed from the detailed numerical analysis [39.39, 44]. Thus, the effective Hamiltonian constraint is, for $N = 1$

$$C_{\text{eff}} = \frac{3}{8\pi G\gamma^2}\frac{\sin^2(\lambda b)}{\lambda^2}V - C_{\text{matt}}\,, \tag{39.25}$$

which leads to modified Friedman and Raychaudhuri equations on computing the Hamilton's equations of motion (as we shall see below). Using (39.25) one can find that the energy density $\rho = H_{\text{matt}}/V$ equals $3\sin^2(\lambda b)/(8\pi G\gamma^2\lambda^2)$. Since the latter reaches its higher possible value when $\sin^2(\lambda b) = 1$, the density has a maximum given by

$$\rho_{\max} = \frac{3}{8\pi G\gamma^2\lambda^2}\,. \tag{39.26}$$

Thus, we see that that the maximum energy density obtained from the effective Hamiltonian is identical to the supremum ρ_{sup} for the density operator in $k = 0$, LQC. The difference is, of course, that in the effective dynamics every trajectory undergoes a bounce and reaches the maximum possible density, while in the quantum theory not every state is close to the critical density at the quantum bounce.

It is easy to solve for the dynamics defined by the effective Hamiltonian. The equations of motion are found using the effective constraint: $\partial_t F =: \dot F = \{F, C\}$, with t the cosmic time. The only equation of motion different from the classical one (on the constraint surface) is

$$\dot V = \frac{3}{\gamma\lambda}V\sin(\lambda b)\cos(\lambda b)\,, \tag{39.27}$$

leading to the modified Friedman equation for the Hubble parameter

$$H^2 := \left(\frac{\dot a}{a}\right)^2 = \left(\frac{\dot V}{3V}\right)^2 = \frac{8\pi G}{3}\rho\left(1 - \frac{\rho}{\rho_{\max}}\right), \tag{39.28}$$

where $\rho_{\max} = 9/(2\alpha^2)(1/\lambda^2)$ is the scalar field density at the bounce. For every trajectory there are quantum turning points at $b = \pm\pi/(2\lambda)$, where $\dot V = 0$, corresponding to a bounce. Note that, at the bounce $\ddot V|_{\beta=\pi/(2\lambda)} = 2\alpha^2 V\rho_{\max} > 0$, so the bounce corresponds to a minimum of volume. Also, note that the Hubble parameter is absolutely bounded $|H| \leq 1/(2\lambda\gamma)$, indicating that the congruence of cosmological observers can never have caustics, independently of the matter content.

In the case of effective theories the proper time appears as a natural choice for an evolution parameter, but one can always look for internal, relational notions of time. Since, $\dot b \leq 0$ one can choose b as a relational time in the effective theories, and consider the evolution with respect to b. The advantage of this election is that no external time variable is needed. Every trajectory, that corresponds to $b > 0$, has a bounce at $b = \pi/(2\lambda)$, and this value tends to infinity as $\lambda \to 0$.

In the effective theories, we consider the interval $b \in [-\pi/(2\lambda), \pi/(2\lambda))$. One should note that all functions and observables in $\bar\Gamma_\lambda$ are periodic in b with period π/λ. It is then completely equivalent to regard the coordinate b as compactified on a circle. The solutions are defined for every t and are given by [39.60]

$$\cot\lambda b = \frac{3t}{\gamma\lambda}\,, \quad V_\lambda(t) = \frac{\alpha}{3}p_\phi\sqrt{\gamma^2\lambda^2 + 9t^2}\,, \tag{39.29}$$

and

$$\phi_\lambda(t) = \phi_0 + \lambda\varphi + \frac{1}{\alpha}\ln\frac{3t + \sqrt{\gamma^2\lambda^2 + 9t^2}}{3t_0 + \sqrt{\gamma^2\lambda^2 + 9t_0^2}}\,, \tag{39.30}$$

so that $\phi_\lambda(t_0) = \phi_0 + \lambda\varphi$ and the initial condition approaches the classical one (for $t = t_0$) as $\lambda \to 0$. Note that $\phi_\lambda(0) \to (\text{sgn}\, b/\kappa)\ln\lambda$ as $\lambda \to 0$. Let us now consider an intrinsic description of the dynamics in terms of the scalar field. One can solve V as a function of ϕ

$$V_\lambda(\phi) = V_+ e^{\alpha(\text{sgn}b)(\phi - \phi(t_0))} + V_- e^{-\alpha(\text{sgn}b)(\phi - \phi(t_0))}\,, \tag{39.31}$$

where

$$V_+ = \frac{1}{2}\left(V_0 + \sqrt{V_0^2 - \beta^2}\right)$$

and $V_- = \beta^2/4(V_+)^{-1}$, where $V_0 = V(\phi(t_0))$, and $\beta = (1/3)\gamma\lambda\alpha p_\phi$. Note that the effective theory recovers the quantum dynamics of $\langle\hat V\rangle|_\phi$ *exactly*, for all

states of the physical Hilbert space. That is, there are no further quantum corrections to the dynamics of $V_\lambda(\phi)$.

With this, one can see that the effective theory defines an effective homogeneous and isotropic spacetime metric, that takes the form

$$(\mathrm{d}s^2)_{\mathrm{eff}} = -\mathrm{d}t + a^2(t)_{\mathrm{eff}}\,\mathrm{d}\mathbf{x}^2\,, \tag{39.32}$$

with

$$a(t)_{\mathrm{eff}} = \left(\frac{\alpha}{3}\right)^{\frac{1}{3}} \frac{p_\phi}{\mathring{V}} (\gamma^2\lambda^2 + 9t^2)^{\frac{1}{6}}\,.$$

It is trivial to see that in the $\lambda \to 0$ limit, one recovers the classical spacetime metric satisfying Einstein equations.

As we have seen, the quantum corrections captured by the effective Hamiltonian modify the Friedman equation in a nontrivial way, ensuring that quantum effects become important near the Planck scale in such a way that a repulsive force is capable of stopping the collapsing universe and turn it around into an expanding phase. Let us explore a little bit more how this quantum repulsive force can be seen. First, a modified Raychaudhuri equation can be written [39.61]

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\rho\left(1 - 4\frac{\rho}{\rho_{\max}}\right) - 4\pi G P\left(1 - 2\frac{\rho}{\rho_{\max}}\right). \tag{39.33}$$

It is also illustrative to write an equation for the rate of change of the Hubble parameter [39.62]

$$\dot{H} = -4\pi G(\rho + P)\left(1 - 2\frac{\rho}{\rho_{\max}}\right). \tag{39.34}$$

These equations imply that the matter conservation equation

$$\dot{\rho} + 3H(\rho + P) = 0\,, \tag{39.35}$$

has the same form as in the classical theory, even when both Friedman and Raychaudhuri equations suffer loop quantum corrections. From (39.34) one sees that, for matter satisfying the WEC, there is a superinflationary phase, corresponding to $\dot{H} > 0$, whenever the matter density satisfies $\rho > \rho_{\max}/2$. Note that in the $\lambda \to 0$ limit, we recover the corresponding classical equations.

Another system of interest, for the remainder sections of this chapter, is a scalar field subject to a potential $V(\phi)$. Even for the simplest potential $V(\phi) = m^2\phi^2/2$ the classical dynamics is drastically modified; after the big bang there is a, *slow roll*, inflationary period. A pressing question is how this dynamics gets modified in the effective LQC scenario. We know that every trajectory follows the effective Friedman equation (39.28) and has a bounce when $\rho = \rho_{\max}$, followed by a period of superinflation. How does that behavior affect the presumed inflationary period occurring at much smaller densities? First note that in that case, the energy density has the form: $\rho = \dot{\phi}^2/2 + m^2\phi^2/2$, so there is a convenient way of depicting the bounce as the curve, in the $(\phi, \dot{\phi})$ plane, satisfying $\rho_{\max} = \dot{\phi}^2/2 + m^2\phi^2/2$. The dynamics is therefore bounded by such ellipsoid. The equation satisfied by the scalar field has the same form as in the classical case: $\ddot{\phi} + 3H\dot{\phi} + V_{,\phi} = 0$. One can solve these equations numerically [39.63, 64] and finds that after the superinflationary phase, the dynamics follows very closely the GR dynamics and exhibits an *attractor* behavior as well. As we shall see in later sections, this feature of the dynamics is responsible for phenomenologically relevant inflation to be generic.

Let us end this part with some comments:

i) This set of effective equations has the property that one recovers General Relativity in the small density *IR* limit, and that they are independent of the fiducial $\mathcal{V}$. These are nontrivial requirements that impose strong conditions on the particular form of the quantum constraint operator [39.65].
ii) Inverse volume effects can introduce modifications to the effective equations that have various consequences, such as loss of the universal conservation equation for matter, and extra superinflationary corrections. However, the physical validity of considering such inverse correction for the $k = 0$ is seriously challenged.
iii) It has been shown that for generic matter content, the LQC effective equations imply that strong singularities are generically resolved [39.61].
iv) A consistency check for the validity of effective equations pertains to the behavior of appropriately defined semiclassical states. Such states have been constructed and the predictions of the effective theory put to the test [39.42, 43]. It was shown that both the density at the bounce and the minimum value of volume are very well described by the effective theory.

$k = 1$ FLRW

Let us now start with the isotropic closed FLRW model. As discussed before, there are two quantization avail-

able for this model. Correspondingly, the effective equations will yield two inequivalent theories. For the first quantization, based in the curvature as defined by closed holonomies, and neglecting the so-called inverse triad corrections, one can arrive at the form of the effective Hamiltonian constraint

$$\begin{aligned} C_{\text{eff}} = &-\frac{3}{8\pi G\gamma^2\lambda^2} V \\ &\times \left[\sin^2(\lambda b - D) - \sin^2 D + (1+\gamma^2) D^2\right] \\ &+ \rho V \,, \end{aligned} \tag{39.36}$$

with $D := \lambda\vartheta / V^{1/3}$. We can now compute the equations of motion from the effective Hamiltonian as

$$\begin{aligned} \dot{V} = \{V, C_{\text{eff}}\} &= \{V, b\} \frac{\partial C_{\text{eff}}}{\partial b} \\ &= \frac{3}{\lambda\gamma} V \sin(\lambda b - D) \cos(\lambda b - D) \,. \end{aligned}$$

From here, we can find the expansion as

$$\begin{aligned} \theta = \frac{\dot{V}}{V} &= \frac{3}{\lambda\gamma} \sin(\lambda b - D) \cos(\lambda b - D) \\ &= \frac{3}{2\lambda\gamma} \sin 2(\lambda b - D) \,. \end{aligned} \tag{39.37}$$

From the above equation we can see that the Hubble parameter is also absolutely bounded by $|H| = |\theta|/3 \leq 1/2\lambda\gamma$. We can now compute the modified, *effective Friedman equation*, by computing $H^2 = \frac{\theta^2}{9}$

$$\begin{aligned} H^2 = &\frac{1}{\lambda^2\gamma^2} \left(\frac{8\pi G\gamma^2\lambda^2}{3} \rho + \sin^2 D - (1+\gamma^2) D^2 \right) \\ &\times \left(1 - \frac{8\pi G\gamma^2\lambda^2}{3} \rho - \sin^2 D + (1+\gamma^2) D^2 \right) \\ = &\frac{8\pi G}{3} (\rho - \rho_1) \left(1 - \frac{\rho - \rho_1}{\rho_{\max}} \right) , \end{aligned} \tag{39.38}$$

where $\rho_1 = \rho_{\max}[(1+\gamma^2)D^2 - \sin^2 D]$ and $\rho_{\max} = 3/(8\pi G\gamma^2\lambda^2)$ is the *critical density* of the $k = 0$ FLRW model.

Let us now consider the other quantization, based on defining the connection using holonomies along open paths. As mentioned before, this is the only available route for anisotropic cosmologies when there is intrinsic curvature (such as in Bianchi II and IX). The effective Hamiltonian constraint one obtains from that quantum theory [39.46], when neglecting inverse scale factor effects (as was done in [39.44, 66]), is

$$C_{\text{eff}} = -\frac{3}{8\pi G\gamma^2\lambda^2} V \left[(\sin\lambda b - D)^2 + \gamma^2 D^2 \right] + \rho V \,. \tag{39.39}$$

It is then straightforward to compute the corresponding effective equations of motion. In particular, by computing $\dot{V} = \{V, C_{\text{eff}}\}$, we can find the expression for the expansion as

$$\theta = \frac{3}{\lambda\gamma} \cos\lambda b \, (\sin\lambda b - D) \,. \tag{39.40}$$

Note that in this case, the expansion (and Hubble) is not absolutely bounded, due to the presence of the term linear in D. An important feature of these effective equations is that they describe with great accuracy the expectation value of volume during the numerical evolution of semiclassical quantum states [39.44]. It is also worth to note that for large values of the recollapse volume, the effective and the classical equations coincide. In the case of the connection-based quantization [39.46], there are two different bounces, that approach the unique bounce of the curvature-based equations when the universe grows to be large [39.46]. Let us now consider the effective equations for anisotropic models.

Anisotropic Models: Bianchi I, II, and IX

Considering the effective description of anisotropic models is interesting in view of the BKL conjecture [39.67–69], that states that locally, generic spacetimes approaching the classical singularity behave as a combination of Bianchi cosmological models. The effective Hamiltonian constraint for Bianchi I and II can be written in a single expression [39.52, 55, 70],

$$\begin{aligned} C_{\text{BII}} &= \frac{p_1 p_2 p_3}{8\pi G\gamma^2\lambda^2} [\sin\bar{\mu}_1 c_1 \sin\bar{\mu}_2 c_2 + \sin\bar{\mu}_2 c_2 \sin\bar{\mu}_3 c_3 \\ &\qquad + \sin\bar{\mu}_3 c_3 \sin\bar{\mu}_1 c_1] \\ &\quad + \frac{1}{8\pi G\gamma^2} \\ &\qquad \times \left[\frac{\alpha (p_2 p_3)^{3/2}}{\lambda\sqrt{p_1}} \sin\bar{\mu}_1 c_1 - (1+\gamma^2) \left(\frac{\varepsilon p_2 p_3}{2 p_1} \right)^2 \right] \\ &\quad - \frac{p_\phi^2}{2} \approx 0 \end{aligned}$$

where the parameter ε allows us to distinguish between Bianchi I ($\varepsilon = 0$) and Bianchi II ($\varepsilon = 1$). This Hamiltonian together with the Poisson brackets $\{c^i, p_j\} = 8\pi G\gamma\delta^i_j$ and $\{\phi, p_\phi\} = 1$ gives the effective equations of motion. In these previous effective Hamiltonians, one chooses the lapse $N = V$.

In Bianchi IX, one chooses $N = 1$ to include more inverse triad corrections. Then the effective Hamiltonian looks like that of the Bianchi I plus some extra terms that capture the information about the intrinsic curvature [39.70].

Let us now discuss the issue of singularity resolution when these equations are studied numerically. The main features that these systems posses can be summarized as follows:

1. All solutions have a bounce. In other words, singularities are resolved. In the closed FRW and the Bianchi IX model, there are infinite number of bounces and recollapses due to the compactness of the spatial manifold.
2. One can have a different kind of bounce dominated by shear σ, but only in Bianchi II and IX. In Bianchi I, the dynamical contribution from matter is always bigger than the one from the shear, even in the solution which reaches the maximal shear at the bounce [39.71].
3. In the flat isotropic model all the solutions to the effective equations have a maximal density equal to the critical density, and a maximal expansion ($\theta^2_{\text{max}} = 6\pi G\rho_{\text{max}} = 3/(2\gamma\lambda)$) when $\rho = \rho_{\text{crit}}/2$. For FRW $k = 1$ model, every solution has its maximum density but in general the density is not absolutely bounded. In the effective theory which comes from connection-based quantization, expansion can tend to infinity. For the other case, expansion has the same bound as the flat FRW model. However, by adding some more corrections from inverse triad term, one can show that actually in both effective theories the density and the expansion have finite values.
4. For Bianchi I, in all the solutions ρ and θ are upperly bounded by its values in the isotropic case and σ is bounded by $\sigma^2_{\text{max}} = 10.125/(3\gamma^2\lambda^2)$ [39.72–74]. For Bianchi II, θ, σ and ρ are also bounded, but for larger values than the ones in Bianchi I, i. e., there are solutions where the matter density is larger than the critical density. With point-like and cigar-like classical singularities [39.71], the density can achieve the maximal value ($\rho \approx 0.54\rho_{\text{Pl}}$) as a consequence of the shear being zero at the bounce and curvature different from zero.
5. For Bianchi IX the behavior is the same as in closed FRW, if the inverse triad corrections are not used, then the geometric scalars are not absolutely bounded. But if the inverse triad corrections are used then, on each solution, the geometric scalars are bounded but there is not an absolute bound for all the solutions [39.70, 74].

39.3 Inhomogeneous Perturbations in LQC

The theory of quantized fields in curved spacetimes has become an essential tool in modern early-universe cosmology. In that framework, one studies the behavior of quantum fields propagating in spacetimes with generic Lorentzian geometries, as in General Relativity. One expects this theory to describe accurately physical processes in situations where we are confident about the validity of its building blocks: a description of matter fields in terms of quantum field theories, and a spacetime geometry given by a smooth, classical spacetime metric. These assumptions are reasonable, for instance, during the inflationary era in which the energy density and curvature are believed to be more than 10 orders of magnitude below the Planck scale. However, earlier in the history of the universe, closer to the Planck era, quantum gravity effects become important and the description of spacetime geometry in terms of a smooth metric is expected to fail. To include physics in the Planck regime QFT in curved backgrounds needs to be generalized to a QFT in *quantum* spacetimes. The singularity-free quantum geometry provided by LQC, summarized in the previous section, provides a suitable arena to formulate such a theory, and the quantization of scalar fields on those quantum cosmologies was introduced by Ashtekar et al. [39.4], and further developed in [39.5, 75–77]. Having in mind the most interesting application of this framework, we summarize here the construction of the QFT of scalar and tensor metric perturbations propagating in a quantum FLRW universe, i. e., the *quantum gravity theory of cosmological perturbations*. For details, see [39.4, 5].

As mentioned in the introduction of this chapter, the construction will follow the guiding principle that has been useful in the quantization of the background: first carry out a truncation of the classical theory to select the sector of General Relativity of interest, and then move to the quantum theory by using LQG techniques. Starting from General Relativity with a scalar field as matter source, we will truncate the phase space to the sector containing cosmological backgrounds *plus* inhomogeneous, gauge invariant, first-order perturbations, and then write down the dynamical equations on that classical, reduced phase space. The main approximation behind this truncation, and underlaying the subsequent quantization, is that the back-reaction of inhomogeneous perturbations on the homogeneous degrees of freedom is neglected. The second step is to move to the quantum theory. Physical states will depend on background homogeneous degrees of freedom as well as on inhomogeneous ones. Our basic approximation, however, enables us to write these quantum states as tensor product of the homogeneous part, which will evolve independently of perturbations, and first-order inhomogeneities thereon. The homogeneous part will therefore be the same as the quantum geometries obtained in the previous section, in which the big-bang singularity is replaced by a bounce. The surprising result appears in the evolution of perturbations. Without further approximation, the evolution of inhomogeneities on those quantum geometries turns out to be *mathematically equivalent* to the quantum theory of those fields propagating on a *smooth* background characterized by a metric tensor. The components of that smooth metric, however, do not satisfy the classical Einstein equation. They are obtained from expectation values of certain combinations of background operators, and incorporate *all* the information of the underlying quantum geometry that is *seen* by perturbations. The message is that the propagation of inhomogeneous perturbations is not sensitive to all the details of the quantum spacetime, but only to certain aspects, which appear precisely in a way that allows one to encode them in a smooth background metric. This is an unforeseen simplification that facilitates enormously the treatment of field theoretical issues.

The last step is to develop the necessary tools to check the self-consistency of this construction. It is necessary to show that, in the physical situations under consideration, the Hilbert space of physical interest contains a large enough subspace in which the back-reaction of perturbations on the background is indeed negligible, in such a way that our initial truncation is justified. That should be done by comparing the expectation value of the Hamiltonian and stress-energy tensor for perturbations with that of background fields. Those computation will require techniques of regularization and renormalization.

39.3.1 The Classical Framework

The goal of this subsection is to summarize the construction of the truncated theory of classical FLRW spacetimes coupled to a scalar field, plus gauge invariant, linear perturbations on it, and write down the equations describing their dynamics. The reader is referred to the extensive literature for more details (see, for instance, [39.78]). We adopt here the Hamiltonian framework which, as shown in [39.79], is particularly transparent on the task of finding gauge invariant variables. It will also provide the appropriate arena to pass to the quantum theory in the next section. For simplicity and for physical interest, we work here with a spatially flat FLRW universe.

The procedure can be divided in three steps:

1. Starting from the full phase space, expand the configuration variables and their conjugate momenta in perturbations, and truncate the expansion at first order. Expand also the constraints of the theory (the scalar and vector constraints) and keep only terms containing zero and first-order perturbations.
2. Use the constraints linear in first-order perturbations to find gauge invariant variables. Those variables coordinatize the so-called truncated reduced phase space.
3. Use the part of the constraints quadratic in zero and first-order perturbations to write down the dynamics.

See [39.5] for further details and subtle points of this construction.

The Truncated Phase Space

Let us consider General Relativity coupled to a scalar field on a spacetime manifold $M = \Sigma \times \mathbb{R}$, with $\Sigma = \mathbb{R}^3$. Due to the infinite volume in Σ, spatial integrals of homogeneous quantities will introduce infrared divergences. To be able to write meaningful mathematical expression, it is convenient to introduce a fiducial cell $\mathcal{V}$ and restrict all integrals to it. $\mathcal{V}$ can be chosen to be arbitrarily large, or at least larger than the observable universe. At the quantum level this will be equivalent to restrict to $\mathcal{V}$ the support of test functions in operator-valued distributions.

To facilitate comparison with the literature on cosmological perturbations, we will work with ADM variables for the gravitational sector, where the canonical conjugated pairs consist in a positive definite 3-metric on Σ, q_{ab}, and its conjugate momentum p^{ab} (the same analysis can be done in connection variables, by including the corresponding Gauss constraint; see [39.5, 80–85]). The full phase space Γ consists of quadruples $\{q_{ab}(\boldsymbol{x}), p^{ab}(\boldsymbol{x}), \Phi(\boldsymbol{x}), \Pi(\boldsymbol{x})\} \in \Gamma$, where $\Pi(\boldsymbol{x})$ is the conjugate momentum of the scalar field $\Phi(\boldsymbol{x})$. Because we are interested in expanding around $\Gamma_{\text{hom}} \subset \Gamma$, the (FLRW) isotropic and homogenous sector of Γ, it is convenient to introduce a fiducial flat metric $\mathring{q}_{ab}$, and use it to raise and lower indices. We will denote $\boldsymbol{x} = (x_1, x_2, x_3)$ the Cartesian coordinates defined by $\mathring{q}_{ab}$ on $\mathcal{V}$, $\mathring{V}$ the volume of $\mathcal{V}$ with respect to $\mathring{q}_{ab}$, which we take equal to one to simplify the notation, and $\mathring{q} = 1$ the determinant of $\mathring{q}_{ab}$.

Consider now curves $\gamma[\epsilon]$ in Γ, which pass through Γ_{hom} at $\epsilon = 0$. Expanding the phase space variables around $\epsilon = 0$, we have

$$\begin{aligned} q_{ab}[\epsilon](\boldsymbol{x}) &= a^2 \mathring{q}_{ab} + \epsilon \delta q_{ab}^{(1)}(\boldsymbol{x}) + \cdots + \frac{\epsilon^n}{n!} \delta q_{ab}^{(n)}(\boldsymbol{x}) \\ &\quad + \ldots \\ \Phi[\epsilon](\boldsymbol{x}) &= \phi + \epsilon \varphi^{(1)}(\boldsymbol{x}) + \ldots , \end{aligned} \tag{39.41}$$

and similarly for the conjugated momenta. It is convenient to consider the first-order perturbations $\delta q_{ab}^{(1)}(\boldsymbol{x}), \delta p^{ab(1)}(\boldsymbol{x}), \varphi^{(1)}(\boldsymbol{x}), \pi^{(1)}(\boldsymbol{x})$ as *purely inhomogeneous* functions of $\boldsymbol{x}$, in the sense that the integral of any of them on $\mathcal{V}$ is zero. By truncating the above expansions at first order we obtain the *truncated* phase space, made of four pairs of conjugate variables

$$\begin{aligned} \Gamma_{\text{Trun}} &= \left\{ \left(a, P_a, \phi, p_{(\phi)}, \delta q_{ab}^{(1)}, \delta p^{ab(1)}, \varphi^{(1)}, \pi^{(1)} \right) \right\} \\ &= \Gamma_{\text{hom}} \times \Gamma_1 . \end{aligned}$$

From now on, we will work only with first-order perturbations, so we will omit the superindex (1) to simplify the notation.

Because of the homogeneity of the background it is convenient to Fourier transform the perturbation fields and carry out the standard scalar–vector-tensor decomposition, in which the six degrees of freedom of δq_{ab} are decompose into two scalar, two vector, and two tensor modes (see, e.g., [39.5, 79] for details). Because perturbations are inhomogeneous, the restriction to the fiducial cell $\mathcal{V}$ is not strictly necessary, and one can avoid the artificial quantization of $\boldsymbol{k}$ that it introduces. However, from the physical point of view one can absorb modes with wavelength larger than the observable universe in the background. Therefore, we will consider that the Fourier integrals incorporate an infrared cut-off k_0 provided by the size of the observable universe.

Constraints and Reduced Phase Space

A similar expansion to (39.41) can be carried out for the constraints. In General Relativity, the Hamiltonian is a sum of constrains, the familiar scalar $\mathbb{S}[N]$, and vector $\mathbb{V}[\boldsymbol{N}]$ constraints. If $\gamma[\epsilon]$ is now a curve that lies in the constraint hyersurface of Γ, and intersects Γ_{hom} at $\epsilon = 0$, by referring to the constraints collectively as $C(q^{ab}, p_{ab}, \Phi, \Pi)$ (suppressing the smearing fields for simplicity), we expand around $\epsilon = 0$ to obtain a hierarchy of equations

$$\begin{aligned} &C^{(0)} := C|_{\epsilon=0} = 0 \,, \quad C^{(1)} := \frac{dC}{d\epsilon}|_{\epsilon=0} = 0 \,, \quad \ldots \\ &C^{(n)} := \frac{d^n C}{d\epsilon^n}|_{\epsilon=0} = 0 \,, \quad \ldots \end{aligned} \tag{39.42}$$

- The zeroth-order constraint, $C^{(0)} = 0$, is just the restriction of the full constraint to the homogeneous subspace Γ_{hom}. The zeroth-order vector constraint is trivially satisfied because of the gauge fixing on the zeroth-order variables, introduced by the use of the fiducial metric $\mathring{q}_{ab}$ in (39.41). The zeroth-order scalar constraint $\mathbb{S}_0$ is quadratic in zeroth-order variables and can be interpreted as the generator of background dynamics. This dynamics is exactly the same as that of the unperturbed theory.
- First-order constraints are linear in first-order variables. They generate gauge transformations in Γ_{Trun} and, as usual, tell us that some of our degrees of freedom are not physical. Initially we have $6(\times\infty)$ degrees of freedoms in $\delta q_{ab}(\boldsymbol{x})$, plus one degree of freedom in the scalar field $\varphi(\boldsymbol{x})$, a total of 7. As mentioned above, $\delta q_{ab}(\boldsymbol{x})$ is conveniently decomposed in Fourier space into two scalars, two vector, and two tensor modes. We have the scalar and three vector constraints, a total of 4. Therefore, the number of physical degrees of freedom is $7 - 4 = 3$. There is an elegant systematic procedure to construct gauge invariant variables out of those three degrees of freedom, and we refer the reader to [39.79] for details. It can be summarize as follows. In FLRW backgrounds, scalar perturba-

tions are affected by the scalar constraint and only one of the vector constraints; they reduce the three scalar degrees of freedom that we have initially, two from gravity and one from the matter sector, to only one physical scalar mode. Vector perturbations are affected by two of the vector constraints that kill completely the vector modes. In other words, in the absence of matter with vector degrees of freedom, as in the case we are studying, there are no physical vector perturbations. Tensor modes are not affected by any of the constraints and therefore the two original tensor modes are the physical ones, i. e., they are gauge invariant. In summary, after imposing the constraints we are left with one scalar degree of freedom, which we choose to be the familiar Mukhanov variable $\mathcal{Q}$, and two tensor modes $\mathcal{T}^{(1)}$ and $\mathcal{T}^{(2)}$. They are gauge invariant variables. and together with their conjugate momenta form the *reduced*, truncated phase space of first-order perturbations, $\tilde{\Gamma}_{\mathrm{Trun}}$. Equations $C^{(n)}=0$ with $n>1$ do not add further constraints on first-order perturbations.

- The second-order constraints in the full phase space Γ involve terms quadratic in first-order perturbations as well as linear terms in second-order perturbations. When a second-order constraint $C^{(2)}$ is restricted to the truncated phase space $\tilde{\Gamma}_{\mathrm{Trun}}$, terms containing second-order perturbations are disregarded, and the resulting combination of quadratic terms in first-order perturbation with coefficients containing background quantities, $\tilde{C}^{(2)}$, *is no longer a constraint*. The truncated second-order scalar constraint $\tilde{\mathbb{S}}_2$ is interpreted as the Hamiltonian that generates the dynamics of gauge invariant first-order perturbations. It has the form

$$\tilde{\mathbb{S}}_2=\tilde{\mathbb{S}}_2^{(\mathcal{Q})}+\tilde{\mathbb{S}}_2^{(\mathcal{T}^{(1)})}+\tilde{\mathbb{S}}_2^{(\mathcal{T}^{(2)})}\,,$$

which indicates that scalar and tensor modes evolve independently of each other, where

$$\tilde{\mathbb{S}}_2^{(\mathcal{T})}[N]=\frac{N}{2(2\pi)^3}\times\int \mathrm{d}^3k\left(\frac{4\kappa}{a^3}|\mathfrak{p}_{\boldsymbol{k}}^{(\mathcal{T})}|^2+\frac{ak^2}{4\kappa}|\mathcal{T}_{\boldsymbol{k}}|^2\right). \tag{39.43}$$

with $\kappa=8\pi G$. The two tensor modes behave identically, and we have denoted them collectively by $\mathcal{T}$. For pedagogical reasons, we only write down the expressions for tensor perturbations. See [39.5, 86] for explicit expressions for scalar modes. In the above equations $\mathfrak{p}_{\boldsymbol{k}}^{(\mathcal{T})}$ are the conjugate momenta of $\mathcal{T}_{\boldsymbol{k}}$, with Poisson brackets

$$\left\{\mathcal{T}_{\boldsymbol{k}},\mathfrak{p}_{-\boldsymbol{k}'}^{(\mathcal{T})}\right\}=(2\pi)^3\delta(\boldsymbol{k}-\boldsymbol{k}')\,.$$

Tensor perturbations, except for the constant factor $1/(2\sqrt{\kappa})$ that provides the appropriate dimensions, behave exactly as massless, free scalar fields (scalar perturbations $\mathcal{Q}_{\boldsymbol{k}}$ behave as a scalar field subject to a time dependent *emergent* potential). The (homogeneous) lapse function N indicates the time coordinate one is using. For instance, $N=1$ corresponds to standard cosmic time t, $N=a$ to conformal time η, and $N=a^3/p_{(\phi)}$ to choosing the scalar field ϕ as a time variable, which turns out to be the natural choice in the quantum theory.

To summarize, the phase space of physical interest is the reduced, truncated phase space $\tilde{\Gamma}_{\mathrm{Trun}}$ made of elements

$$\Big\{(a,P_a,\phi,p_{(\phi)});\;\left(\mathcal{Q}_{\boldsymbol{k}},\mathfrak{p}_{\boldsymbol{k}}^{(\mathcal{Q})},\mathcal{T}_{\boldsymbol{k}}^{(1)},\mathfrak{p}_{\boldsymbol{k}}^{(\mathcal{T}^{(1)})},\mathcal{T}_{\boldsymbol{k}}^{(2)},\mathfrak{p}_{\boldsymbol{k}}^{(\mathcal{T}^{(2)})}\right)\Big\}\in\tilde{\Gamma}_{\mathrm{Trun}}\,.$$

The homogenous degrees of freedom evolve with the zeroth-order Hamiltonian. This evolution takes place entirely in Γ_{hom}, and is independent of perturbations, reflecting the main approximation of the truncated theory. The homogenous dynamical trajectory can then be *lifted* to $\tilde{\Gamma}_{\mathrm{Trun}}$, providing a well-defined evolution of first-order perturbations on the homogenous background. This evolution is specified by the Hamiltonian $\tilde{\mathbb{S}}_2$.

39.3.2 Quantum Theory of Cosmological Perturbations on a Quantum FLRW

Quantization of $\tilde{\Gamma}_{\mathrm{Trunc}}$

In this section, we pass to the quantum theory starting from the reduced, truncated phase space $\tilde{\Gamma}_{\mathrm{Trun}}$. The structure of the classical phase space $\tilde{\Gamma}_{\mathrm{Trun}}=\Gamma_{\mathrm{hom}}\times\tilde{\Gamma}_1$ suggests that in the quantum theory the total wave function Ψ has the form

$$\Psi(a,\mathcal{T}_{\boldsymbol{k}},\phi)=\Psi_0(a,\phi)\otimes\psi(\mathcal{T}_{\boldsymbol{k}},\phi)\,. \tag{39.44}$$

This product structure is maintained as long as the test field approximation holds. Because back-reaction is neglected, the background part Ψ_0 evolves independently of perturbations, and the solutions for Ψ_0

are the ones obtained in Sect. 39.2. When written in terms of the relational time ϕ, they satisfy the equation $\hat{p}_{(\phi)}\Psi_0 \equiv -\mathrm{i}\hbar\partial_\phi \Psi_0 = \hat{H}_0\Psi_0$, where the operator $\hat{H}_0 \equiv \sqrt{\Theta}$ is obtained from expressions (39.3) and (39.11). The remaining task is to *lift* this trajectory to the full Hilbert space, by writing down the quantum theory for ψ propagating on the quantum geometry specified by Ψ_0. The evolution of ψ will be specified by the operator analogue of $\tilde{\mathbb{S}}_2^{(\mathcal{T})}$, which generates the dynamics in the classical phase space. In the classical theory $\tilde{\mathbb{S}}_2^{(\mathcal{T})}$ depends not only on inhomogeneous degrees of freedoms, but also on the homogeneous ones via the scale factor a. Therefore, in the quantum theory the corresponding operator will act on perturbations ψ as well as on Ψ_0.

Our goal is to generalize the theory of QFT in curved spacetimes in which, on the one hand, quantum fields propagate in an *evolving* classical FLRW specified by $a_{\mathrm{cl}}(\eta)$ and, on the other hand, perturbations are commonly quantized using the Heisenberg picture. Therefore, to facilitate the comparison, we pass in this section to the Heisenberg picture. In obtaining the evolution equations for the operator $\hat{\mathcal{T}}_k$ and its conjugated momentum we will use ϕ as internal time, because it is the evolution variable that appears naturally in the quantum theory, while standard cosmic or conformal time are represented by operators. Internal time ϕ corresponds to use the lapse function $N = a^3/p_{(\phi)}$ in the expression (39.43). By choosing an appropriate factor ordering to convert it to an operator, we have (as it is common in quantum theory, we are not free of factor ordering ambiguities)

$$\begin{aligned}\partial_\phi \hat{\mathcal{T}}_k(\phi) &= \frac{\mathrm{i}}{\hbar}\left[\hat{\mathcal{T}}_k, \hat{\mathbb{S}}_2^{(\mathcal{T})}\right] = 4\kappa\left(\hat{p}_{(\phi)}^{-1} \otimes \hat{\mathfrak{p}}_k^{(\mathcal{T})}\right);\\ \partial_\phi \hat{\mathfrak{p}}_k^{(\mathcal{T})}(\phi) &= \frac{\mathrm{i}}{\hbar}\left[\hat{\mathfrak{p}}_k^{(\mathcal{T})}, \hat{\mathbb{S}}_2^{(\mathcal{T})}\right]\\ &= -\frac{k^2}{4\kappa}\left(\hat{p}_{(\phi)}^{-1/2}\hat{a}^4(\phi)\hat{p}_{(\phi)}^{-1/2} \otimes \hat{\mathcal{T}}_k\right).\end{aligned} \tag{39.45}$$

These equations involve background operators as well as perturbations. However, the test field approximation allows us to *trace over* the background degrees of freedom. This can be done by taking expectation value with respect to the background wave function Ψ_0 (in the Heisenberg picture) obtained in the previous section

$$\begin{aligned}\partial_\phi \hat{\mathcal{T}}_k &= 4\kappa\left\langle \hat{H}_0^{-1}\right\rangle \hat{\mathfrak{p}}_k^{(\mathcal{T})},\\ \partial_\phi \hat{\mathfrak{p}}_k^{(\mathcal{T})} &= -\frac{k^2}{4\kappa}\left\langle \hat{H}_0^{-1/2}\hat{a}^4(\phi)\hat{H}_0^{-1/2}\right\rangle \hat{\mathcal{T}}_k,\end{aligned} \tag{39.46}$$

where background operators have been replaced by expectation values and, additionally, we have used the evolution equation $\hat{p}_{(\phi)}\Psi_0 = \hat{H}_0\Psi_0$. The test field approximation ensures that we are not losing any information when passing from (39.45) to (39.46). These are the Heisenberg equations for perturbations, in which the coefficients are given by *expectation values of background operators in the quantum geometry specified by* Ψ_0. This is a quantum field theory of cosmological perturbation on a *quantum* FLRW universe. Note that the above equation is exact, and not further approximation has been made beyond the test field approximation.

In this theory, spacetime geometry is not described by a unique classical metric, it is rather characterized by a probability distribution Ψ_0 that contains the unavoidable quantum fluctuations. The propagation of perturbations is sensitive to those fluctuations. However, it is remarkable that those effects can be encoded in a couple of expectation values of background operators: $\langle \hat{H}_0^{-1}\rangle$ and [39.4, 5]

$$\langle \hat{H}_0^{-\frac{1}{2}}\hat{a}^4(\phi)\hat{H}_0^{-\frac{1}{2}}\rangle .$$

Borrowing the analogy from [39.5], this is similar to what happens in the propagation of light in a medium: the electromagnetic waves interact in a complex way with the atoms in the medium, but the net effect of those interactions can be codified in a few parameters, such as the refractive index. Similarly, although the final equations (39.46) depend in a simple way on the quantum geometry, it had been very difficult to guess the precise *moments* of the quantum geometry that are involved in the evolution of perturbations.

We can now compare the above evolution equations with the familiar quantum field theory of cosmological perturbations on classical FLRW geometries, in which the Heisenberg equations, when ϕ is used as time, are written in terms of the classical background quantities $a(\phi)$ and $p_{(\phi)}$ as

$$\begin{aligned}\partial_\phi \hat{\mathcal{T}}_k &= \frac{4\kappa}{p_{(\phi)}}\hat{\mathfrak{p}}_k^{(\mathcal{T})};\\ \partial_\phi \hat{\mathfrak{p}}_k^{(\mathcal{T})} &= -\frac{k^2}{4\kappa}\frac{a(\phi)^4}{p_{(\phi)}}\hat{\mathcal{T}}_k .\end{aligned} \tag{39.47}$$

Comparing with (39.46) we see that the QFT in a quantum background Ψ_0 is *indistinguishable* from a QFT on

a *smooth FLRW metric*

$$\tilde{g}_{ab}\,\mathrm{d}x^a\,\mathrm{d}x^b \equiv \mathrm{d}\tilde{s}^2 = -(\tilde{p}_{(\phi)})^{-2}\tilde{a}^6(\phi)\,\mathrm{d}\phi^2 + \tilde{a}(\phi)^2\,\mathrm{d}\mathbf{x}^2\,, \tag{39.48}$$

where

$$(\tilde{p}_{(\phi)})^{-1} = \langle \hat{H}_0^{-1}\rangle \quad \text{and} \quad \tilde{a}^4 = \frac{\left\langle \hat{H}_0^{-\frac{1}{2}}\hat{a}^4(\phi)\hat{H}_0^{-\frac{1}{2}}\right\rangle}{\langle \hat{H}_0^{-1}\rangle}\,. \tag{39.49}$$

In terms of the more familiar conformal time used in cosmology, we have $\mathrm{d}\tilde{s}^2 = \tilde{a}^2(\tilde{\eta})(-\mathrm{d}\tilde{\eta}^2 + \mathrm{d}\boldsymbol{x}^2)$, with $\mathrm{d}\tilde{\eta} = [\tilde{a}^2(\phi)]\tilde{p}_{(\phi)}^{-1}\,\mathrm{d}\phi$. This smooth metric captures all the information of quantum geometry that is *seen* by perturbations. Note that its components contain $\hbar$ and it does not satisfy the Einstein equation, not even the LQC effective equations.

In terms of this smooth metric, we can write the Heisenberg equations (39.46) as a second-order differential equation

$$\hat{\mathcal{T}}_{\boldsymbol{k}}'' + 2\frac{\tilde{a}'}{\tilde{a}}\hat{\mathcal{T}}_{\boldsymbol{k}}' + k^2\hat{\mathcal{T}}_{\boldsymbol{k}} = 0\,, \tag{39.50}$$

where the prime now denotes derivative with respect to $\tilde{\eta}$. This equation is mathematically equivalent to the familiar formulation of QFT in classical FLRW spacetime, where all the effects of the quantum background geometry have been encoded in a *dressed, smooth metric tensor* $\tilde{g}_{ab}$. This unexpected mathematical analogy highly simplifies the analysis, not only conceptually, but also at the technical level. It allows to extend well-established techniques from classical spacetimes to define the physical Hilbert space and the appropriate regularization and renormalization of composite operators on it (see [39.5, 86] for details of that construction). These are the necessary tools to make sense of the momentum integrals appearing in, e.g. the Hamiltonian $\hat{\hat{\mathbb{S}}}_2$, that so far were formal, and to regularize the expectation value of the energy-momentum tensor in the physical Hilbert space.

Criterion for Self-Consistency

The last step in the construction is to check whether the underlaying approximation in our truncated theory, the test field approximation, is satisfied throughout the evolution. In our QFT in quantum spacetimes this question translates to check whether the expectation value of the stress-energy tensor can be neglected when compared to the background one. However, in an homogeneous and isotropic background a sufficient condition for this to be satisfied is that energy density on scalar and tensor perturbations $\langle\hat{\rho}(\tilde{\eta})\rangle$ be much smaller than the background energy density $\langle\rho_0\rangle$ *at any time* during dynamical phase of interest [39.5]. It is evident that one can always find states for perturbation for which that requirement is not satisfied. Therefore, the relevant question is: is there a sufficiently large subspace of the physical Hilbert space for which the previous condition on the energy density is satisfied? If the answer is in the affirmative then one has a self-consistent approach in which test-field approximation holds. This is a key question to ensure self-consistency, and has to be answered when this framework is applied to a concrete physical problem, as we do in the next section.

39.3.3 Comments

The previous framework is suitable to face interesting conceptual questions arising in quantum gravity. For instance, when does standard QFT in curved spacetimes become a good approximation? Is it safe to use standard QFT during inflation? This question can be answered straightforwardly because both theories have been written in the same form. From (39.50) it is clear that the standard QFT is recovered in the regime in which the quantum aspects of the geometry can be neglected, and Sect. 39.2.3 provided the conditions under which this happens. When the background energy density $\langle\rho_0\rangle$ is below one thousandth of $\rho_{P\ell}$, quantum corrections become negligible and General Relativity becomes an excellent approximation. This is the regime in which standard QFT arises from the more fundamental framework presented in this section. Therefore, in the inflationary era where $\langle\rho_0\rangle \lesssim 10^{-10}\rho_{P\ell}$, we expect the familiar QFT to be an excellent approximation.

By construction, this framework encompasses the Planck regime and is suitable to discuss trans-Planckian issues and distinguish real problems from apparent ones. In LQG there is a priori no impediment for trans-Planckian modes to exist. It may seem at first that the existence of a minimum area may preclude their existence, but quantum geometry is subtle and, for instance, there is no minimum value for volume or length. In addition, if we pay attention to the construction of the background quantum theory, trans-Planckian quantities appear there without causing problems: the value of the momentum $p_{(\phi)}$ of the background scalar field ϕ

is generally large in Planck units. However, the background energy density is *bounded above* by a fraction of the Planck energy density. Something similar happens in our quantum field theory. There trans-Planckian modes are admitted *as long as the total energy density in perturbations remains small as compared to the background*. That is the real trans-Planckian problem, which becomes a nontrivial issue in the deep Planck regime where the volume of the universe acquires its minimum value.

39.4 LQC Extension of the Inflationary Scenario

The previous sections have summarized the physical ideas and mathematical tools necessary to undertake the quantization of the sector of General Relativity containing the symmetries of cosmological spacetimes and the study of cosmic perturbation thereon. The goal of this section is to apply those techniques to extend the current picture of the evolution of our universe to include the Planck regime.

The cosmological ΛCDM (Lambda-cold dark matter) model with an early phase of inflation contains conceptual limitations that are dictated by the domain of applicability of the physical theories in which it is based: General Relativity and Quantum Field Theory. One needs a theory of quantum gravity to extend the model to include physics at the Planck era. Section 39.4.1 summarizes how, by introducing a scalar field with suitable potential, LQC provides a spacetime in which the big-bang singularity is resolved by the quantum effects of gravity, and in which an inflationary phase arises almost unavoidably at later times. In Sect. 39.4.1, it is shown how the evolution of cosmological perturbation can be extended to include the preinflationary spacetimes provided by LQC. In this sense, the current scenario for the evolution of our universe and the genesis of cosmic inhomogeneities is extended all the way to the big bounce [39.87]. This extension goes beyond the conceptual level, as it appears a narrow window in which the effects of Planck scale physics could be imprinted in the CMB and galaxy distributions, and concrete ideas connecting those effects with forthcoming observations have been proposed.

39.4.1 Inflation in LQC

As we have mentioned in previous sections, after the bounce there is a period of superinflation where $\dot{H} > 0$ until the density reaches half its value at the bounce. It was first hoped that this would be enough to account for the necessary number of *e-foldings*, but this period turns out to be too short when there is no potential for the scalar field. Thus, it is clear that one needs such a potential to compare the LQC predictions with the inflationary paradigm. The simplest case one can consider is quadratic potential $V(\phi) = (1/2)m^2\phi^2$, that has been extensively studied in the literature and is compatible with the 7-years WMAP observations [39.88]. The existence of the bounce solves one of the conceptual challenges that the standard scenario, based on the GR dynamics poses. That is, in the GR dynamics, there is always a past singularity, even in the presence of eternal inflation [39.6]. The standard formalism is therefore, conceptually incomplete.

The question that we shall pose in this part is the following: Can we estimate how probable it is to have enough inflation for the cosmological background? Let us be more precise with the question. We know that every effective trajectory undergoes a bounce, and some of them will experience enough e-foldings and will be of phenomenological relevance. Rather amazingly, WMAP has provided us with a small observational window for the scalar field at the onset of inflation [39.88, 89], written in terms of a reference time t_{k_*} for which a reference mode k_* used by WMAP exited the Hubble radius in the early universe. With an 4.5% accuracy, the data is, in Planck units [39.88, 89]

$$\begin{aligned}\phi(t_{k_*}) &= \pm 3.15\,,\\ \dot{\phi}(t_{k_*}) &= \mp 1.98\times 10^{-7}\,,\\ H(t_{k_*}) &= 7.83\times 10^{-6}\,.\end{aligned}$$

We can now pose the question more precisely. From all the solutions $\mathbb{S}$ to the effective equations in LQC, how many of them pass through the allowed interval? This poses yet another question. How are we going to *count* trajectories? Is there a canonical way of measuring them? A proposal to answer this question was put forward long ago [39.90, 91] based on the idea of using the Liouville measure on phase space $\mathbb{S}$, that is invariant under time evolution. The idea then is to compute the volume of $\mathbb{S}_{\text{wmap}}$, those solutions that pass through

the WMAP window, relative to the total volume of $\mathbb{S}$

$$\text{Prob} = \frac{\text{Vol}(\mathbb{S}_{\text{wmap}})}{\text{Vol}(\mathbb{S})} . \tag{39.51}$$

In order to compute this probability, one has to be careful with the way one measures all possible trajectories (for a discussion see [39.92]).

Let us now rephrase the question that we initially posed at the beginning of this part: What is the relative number of solutions $\tilde{\mathbb{S}}_{\text{wmap}}$ that pass through the observational WMAP window, from the total number of solutions $\tilde{\mathbb{S}}$ at the bounce? As explained before, the probability is computed using formula (39.51), where the volume is now obtained by integrating a uniform distribution (as a function of ϕ). The key to computing the probability is then a detailed knowledge of the global dynamics, for all possible values ϕ_B of the scalar field at the bounce. Extensive numerical evolutions have shown that almost all trajectories fall within the observational window. It is only for the small window $-5.46 < \phi_B < 0.934$ from the total range of $\phi_B \in [-7.44 \times 10^5, 7.44 \times 10^5]$ that the future dynamics lies *outside* the WMAP window [39.89]. For this interval, the probability that the dynamics falls outside of the observational window is *less* than 3×10^{-6}. To understand this, one can see the LQC dynamics as shown in Fig. 39.1, where one considers a uniform distribution at the bounce and follows the dynamics. As can be easily seen, most trajectories funnel into a very small region that is precisely where the WMAP window is. Just before the onset of inflation the density is approximately 10^{-11} smaller than the density at the bounce. At that density the allowed WMAP region is only 4% of the total allowed range in ϕ [39.89]. Thus, as seen in the figure, almost all of the trajectories at the Planck scale fall into a very small region at the onset of inflation [39.64].

One should also note that this attractor feature of the global dynamics explains why the probability is much smaller when computed in General Relativity at the onset of inflation [39.64, 93].

Let us summarize. In LQC it is natural to consider the bounce as the point where to compute probability of inflation. The global dynamics is such that most of the trajectories get funneled into the small WMAP window at the onset of inflation where the density is 11 orders of magnitude smaller than the density at the bounce. Thus, one can conclude that having enough inflation is generic in LQC for the homogeneous and isotropic background.

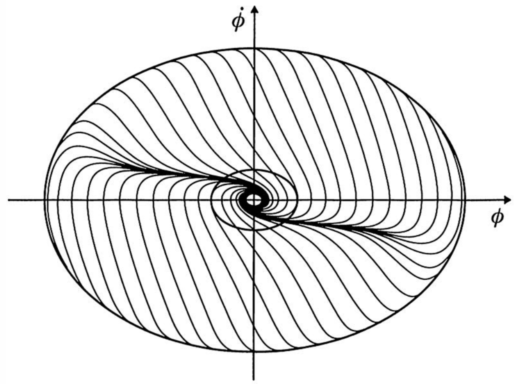


Fig. 39.1 In this figure we plot the exterior, critical density surface $\rho_{\max}$, and a surface of constant density $\rho_{\text{onset}} \ll \rho_{\max}$ (not drawn to scale, of course) on the $(\dot{\phi}, \phi)$ plane. Trajectories with a uniform distribution at the LQC bounce ellipsoid are plotted. Note that trajectories for which there is enough inflation get funneled into a small region in the smaller ρ_{onset} ellipse. Near this surface, the GR and LQC dynamics almost coincide

39.4.2 Preinflationary Evolution of Cosmic Perturbations

In this section, we apply the quantum theory of cosmological perturbations on the quantum, preinflationary spacetime to extend the study of cosmic inhomogeneities all the way back to the Planck era. In addition to the *conceptual* completion provided by the inclusion of Planck scale physics, the resulting framework opens an exciting avenue to extend observations into the Planck regime. Before entering into technical details, we summarize here the physical idea behind this possibility.

It is known since the seminal work by Parker in the 1960s [39.94–96], that a dynamical expansion of the universe is able to excite quanta, or *particles*, of test fields out from an initially vacuum state. This phenomenon of particle creation is one of the main features of QFT in curved spacetimes, and plays a key role in black hole thermal radiance and in the generation of cosmic inhomogeneities during inflation. If $\boldsymbol{k}$ represents a comoving Fourier mode of a test scalar field in FLRW, excitations on that mode may be created if the energy scale provided by the spacetime scalar curvature is comparable to the physical wavelength $\lambda = 2\pi a/k$ at some time during the evolution. The amount of quanta created during a period of expansion in each mode depends on the details of the scale factor $a(t)$ as a function

of time. Let us focus on the finite range of momenta that is accessible in cosmological observations. The previous argument tell us that, even if those modes are *born* in the ground state at time of the bounce, particles may be created during the evolution. The resulting state, e.g. at the onset of inflation, would then depart from the vacuum state at that time as a consequence of the nontrivial evolution, and the spectrum of particles created will carry information about the preinflationary spacetime geometry.

Furthermore, it has been shown in the context of inflation that [39.97–101] the predictions for the CMB and the distribution of galaxies are sensitive to the details of the state describing perturbations at the onset of inflation, and concrete observation have been proposed that could reveal information about that state [39.102–104]. In other words, those observations may reveal information about the propagation of perturbations *before* inflation, when quantum gravity corrections dominate.

In the inflationary scenario observable modes have wavelength much smaller than the radius of curvature at the onset of inflation (in the cosmological argot, modes are deeply inside the Hubble radius). The sometimes implicit assumption in inflationary physics is that, whatever happened before inflation, wavelength of interest were much smaller than the radius of curvature *at any time before inflation*. Under this assumption, preinflationary dynamics for those modes is indistinguishable from an evolution in Minkowski spacetime, and the use of a vacuum state is justified. The relevant question is then: is this assumption accurate in the preinflationary background provided by LQC? More explicitly, consider modes with physical wavelength smaller than the radius of curvature at the beginning of inflation, and propagate them backward in time until the bounce. Do those wavelength generically remain smaller than the radius of curvature of the dressed metric $\tilde{g}_{ab}$ during the entire preinflationary evolution? The detailed analysis of [39.86, 87] shows that the answer to this question is in the negative (Fig. 39.2). While short enough wavelengths (large enough momenta) remain always smaller that the curvature radius, there are modes which at some time during the evolution have physical size comparable to it. The evolution of those modes *is* sensitive to the spacetime curvature and the quantum state at the onset of inflation will depart from the vacuum.

Notice that in LQC the maximum value of the curvature takes place at the bounce time and this value is universal, fixed by the quantum geometry and independent of the form of the scalar field potential. If we call k_R the comoving scale associated with this maximum value of the curvature, we expect excitations with $k \lesssim k_R$ to be created during the evolution, concretely in the Planck regime near the bounce. On the other hand, for modes with $k \gg k_R$ preinflationary dynamics has negligible effect. From this qualitative discussion we may expect observable effects from Planck scale physics in CMB and large scale structure if observations are accessible to modes k around or smaller than the universal scale k_R provided by LQC.

In the remainder of this section, we provide precise computations that support this qualitative physical picture. We start by specifying the initial condition for both background and perturbations at the bounce. We then evolve those perturbations until the end of slow-roll inflation, compute the resulting quantum state and the

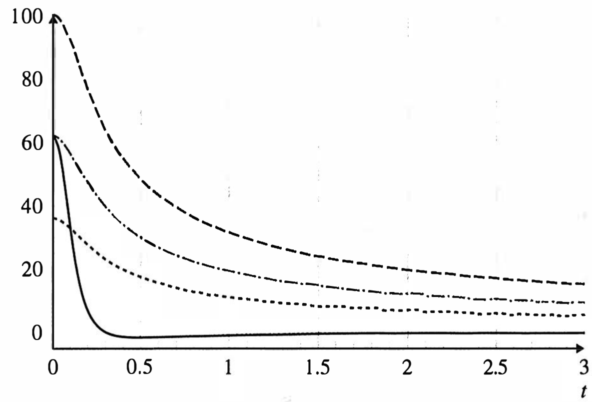


Fig. 39.2 This plot shows: (i) The scalar curvature of the effective geometry (*red solid line*), (ii) The physical momentum squared $(k/\tilde{a}(t))^2$, for $k = 6$ (*dotted black line*), and $k = 10$ (*dashed black line*), and iii) $(k_R/\tilde{a}(t))^2$, where k_R is the comoving scale associated with the maximum value of the curvature (*dotted-dashed green line*); as a function of cosmic time t. By convention, we choose the scale factor of the effective geometry to be one at the bounce, $\tilde{a}(0) = 1$. Both axes are in Planck units. Curvature attains the maximum value at the bounce and decreases very fast after it. Modes with momentum k larger than the scale of curvature at the bounce, $k > k_R$, have physical momentum larger than the curvature during the entire evolution (*dashed black line*). Those modes do not *feel* the curvature and evolve as if they were in Minkowski spacetime. On the other hand, modes that at the bounce have physical momentum smaller that the curvature, $k < k_R$, quickly evolve to become of the same order as the curvature scale (*black dotted line*), and therefore their evolution will differ considerably from that in flat space. At later times those modes also become two energetic to feel the spacetime curvature

power spectrum for scalar and tensor perturbations, and study under what set of initial conditions quantum gravity corrections may be sizeable for observable modes.

Initial Conditions

In the standard inflationary paradigm one specifies *initial data* for the background and perturbations at the onset of slow-roll. From a fundamental point of view, it would be more satisfactory to impose initial conditions at the *beginning* rather than at an intermediate time in the evolution of the universe. In classical cosmology the *beginning* is the big-bang singularity, and it is not possible to unambiguously defined initial condition at that time. In LQC the big bang is replaced by a quantum bounce where physical quantities do not blow up, providing a preferred time to specify initial data.

In the test field approximation, the total wave function naturally decomposes as a product $\Psi = \Psi_0 \otimes \psi$, and this form holds as long as back-reaction of perturbations remains negligible. We need therefore to specify initial data for both, Ψ_0 and ψ.

Background. For computational purposes, it is convenient to make the following further simplification on the background dynamics. As described in Sect. 39.2.1, the background wave function Ψ_0 can be chosen to be highly peaked along the entire evolution, including the deep Planck regime. The *peak* of that wave function describes an effective geometry characterized by the scale factor $\bar{a}(\phi) = \langle \hat{a}(\phi) \rangle$, which satisfies the effective (39.28). Because the dispersion of Ψ_0 remains very small during evolution, it is convenient to ignore quantum fluctuations in our computations, by making a *mean field* approximation in which the expectation values of powers of background operators, such as $\hat{a}$ and $\hat{H}_0$, are replaced by the same powers of their expectation. For instance, in the evolution of quantum inhomogeneities given by (39.50), this is equivalent to replace $\bar{a} \approx \bar{a}$. At the practical level this is an excellent approximation, e.g. numerical errors in simulations turn out to be larger than those introduced by the mean field approximation.

In Sect. 39.4.1 we described the effective preinflationary background arising in LQC for the representative example of a quadratic potential. In that effective geometry initial data is entirely specified by the value of the scalar field at the bounce, ϕ_B, and, unless ϕ_B lies in a small region R around $\phi_B = 0$, the evolution generically finds an inflationary phase at late times compatible with WMAP observations [39.88]. Therefore, we will choose Ψ_0 to be a state sharply peaked in an effective trajectory specified by a value of ϕ_B that lies outside the region R.

The effect of choosing different values of ϕ_B can be understood using the effective equations (39.28) together with numerical simulations. On the one hand, immediately after the bounce the background evolution is entirely dominated by quantum gravity effects, and it is largely insensitive to the concrete value of ϕ_B. Except for very small momenta k, the times at which perturbations $\mathcal{Q}_k$ and $\mathcal{T}_k$ *feel* the spacetime curvature is precisely just after the bounce (Fig. 39.2). Therefore, the features that those modes acquire during the evolution turn out to be quite insensitive to the value of ϕ_B. On the other hand, different values of ϕ_B do modify significantly the spacetime geometry at later times. The larger ϕ_B, the longer it takes to reach the end of slow-roll inflation, or, equivalently, the larger the amount of expansion of the universe between the bounce and the end of slow-roll. A larger amount of expansion implies that observable modes had larger physical momentum at the time of the bounce. Because by convention *we fix the scale factor at the bounce* $\bar{a}_B = 1$ (rather than $\bar{a}_{today} = 1$), the effect of choosing different values of ϕ_B essentially translates into a change in the range of comoving momenta k relevant for observations, moving to larger k's as ϕ_B increases. If $[k_{min}, k_{max} \approx 2000 k_{min}]$ is the window covered by WMAP, we have, for instance, $k_{min} \approx 2.8 \times 10^{-3}$ for $\phi_B = 1$, $k_{min} \approx 0.14$ for $\phi_B = 1.1$ and $k_{min} \approx 8.2$ for $\phi_B = 1.2$. The physical momentum $k/\bar{a}_{today}$ of modes observed today is of course the same in all cases, but the convention $\bar{a}_B = 1$ makes that different amount of expansion (i. e., different ϕ_B) translates into different comoving k for those modes.

Perturbations. As already occurs in classical spacetimes, quantum fields in quantum cosmological backgrounds does not admit a preferred state that we can call *the vacuum*. In backgrounds with large enough number of isometries, e.g. Minkowski or de Sitter spacetime, a preferred ground state can be singled out by imposing symmetry in combination with regularity conditions. In our quantum FLRW we follow the same criteria, and look for quantum states ψ invariant under the isometries of the background, spatial translations and rotations, with appropriate ultraviolet behavior. That selects the family of *fourth adiabatic order vacua* (see [39.5, 86] for further details). As opposed to Poincare or de Sitter invariance, symmetry under spatial translations and rotations is not restrictive enough to select a unique state, but it substantially narrows down the possibilities. This is the set of initial data we choose for perturbations.

The next subsection will summarize the time evolution of different choices of initial state within the family of fourth-order adiabatic vacua, and will show that quantities of interest such as the power spectrum of observable modes, are all very similar.

Physically, the choice of a fourth-order adiabatic vacuum at the time of the bounce corresponds to assume *initial quantum homogeneity*. One is requiring that the portion of the universe corresponding to our observable patch at the time of the bounce is *as homogeneous as quantum mechanics allows*, i. e., only vacuum fluctuation of inhomogeneities are present. This is a strong assumption. The motivation comes from [39.86, 87]:

- In a universe containing a phase of inflation lasting at least for 60 e-folds, the physical size of observable universe was very small at the bounce time, $\lesssim 10\ell_{\mathrm{Pl}}$, for the solutions of interest.
- The *quantum degeneracy force* responsible of the bounce has a diluting effect that may produce homogeneity at scales of the order of the Planck length at the bounce. This is the new ingredient that LQC provides at the time of the bounce to produce homogeneity at Planck scale distances.
- There is a precise sense in which the assumption of quantum homogeneity captures a quantum version of the Weyl curvature hypothesis [39.105].

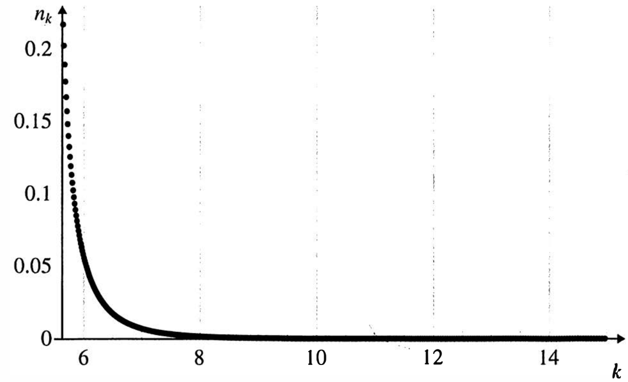


Fig. 39.3 Number n_k of scalar *excitations/particles* with comoving momentum $\boldsymbol{k}$ in the interval $[\boldsymbol{k}, \boldsymbol{k} + \mathrm{d}\boldsymbol{k}]$, per comoving unit volume contained in the evolved state as compared to the BD vacuum during inflation. The plot is computed for $\phi_{\mathrm{B}} = 1.15$ and for the *obvious* fourth-order adiabatic vacuum at the bounce. The *horizontal axis* is in Planck units

Power Spectrum

Our task is to use the equations of the quantum theory summarized in Sect. 39.3.2 to compute the state of cosmic inhomogeneities at the end of the inflationary epoch, by starting from the initial condition specified above for background and perturbations at the time of the bounce.

Due to computational limitations, it is convenient to restrict numerical simulations to backgrounds for which the bounce is kinetic energy dominated, where it has been shown that quantum fluctuations of Ψ_0 remain very small along the entire evolution. Several numerical simulations have been carried out for effective backgrounds with initial conditions $\phi_{\mathrm{B}} \in (0.93, 1.5)$, which turns out to be the most interesting range [39.86]. It is not expected that new features appear for larger values of ϕ_{B}, but computational limitations make difficult to check it explicitly.

For perturbations, simulations have been carried out using different choices of fourth-order adiabatic vacua, and the results are all very similar. Figures 39.3 and 39.4 are obtained by using the *obvious* or *standard* fourth-order vacuum at the bounce time $\tilde{\eta}_{\mathrm{B}}$ (see [39.5] for precise definition), and they show the relevant information of the evolved state.

First of all, to gain intuition on the effect of the preinflationary evolution, we compare the evolved state with the natural vacuum during inflation, the so-called Bunch–Davies (BD) vacuum. Figure 39.3 shows the number n_k of *excitations/particles* with momentum $\boldsymbol{k}$ per comoving unit volume in space and momentum, contained in the evolved state relative to the BD vacuum during inflation. The plot is computed for $\phi_{\mathrm{B}} = 1.15$ but, as explained in Sect. 39.4.2, *Initial Conditions*, it is not altered by choosing a different value inside our family. Changing the value of ϕ_{B} has essentially the effect of shifting the location of the observationally relevant window $[k_{\min}, k_{\max} \approx 2000 k_{\min}]$ in the horizontal axes of the plot, which moves steadily to the right as ϕ_{B} increases. Figure 39.3 is in good agreement with the qualitative arguments presented at the beginning of Sect. 39.4.2. Namely, the preinflationary evolution affects modes with low k, for which a considerable amount of excitations have been *created*. On the contrary, modes with large k remain in the ground state at the onset of inflation. As it was expected, for $k > k_R \approx 7.7$ (recall that k_R is the comoving scale associated with the scalar curvature of the effective metric at the bounce), the number of BD particles contained in the evolved state is very close to zero. Therefore, if $k_{\min} \gtrsim k_R$, that corresponds to $\phi_{\mathrm{B}} \gtrsim 1.2$, the evolved state is indistinguishable from the BD vacuum for observable modes. For $\phi_{\mathrm{B}} \lesssim 1.2$ the state at the onset

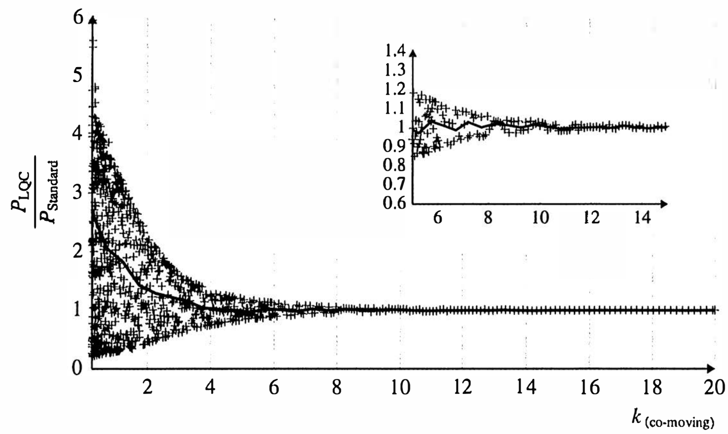


Fig. 39.4 Ratio of the LQC power spectrum for scalar perturbation to the standard inflationary power spectrum. *Crosses* show the ratio for different values of k. The LQC power spectrum oscillates rapidly for small k. The *solid curve* averages over bins of width $\Delta k = 0.5$. The *inset* shows a zoom-in of the interesting region around $k = 9$

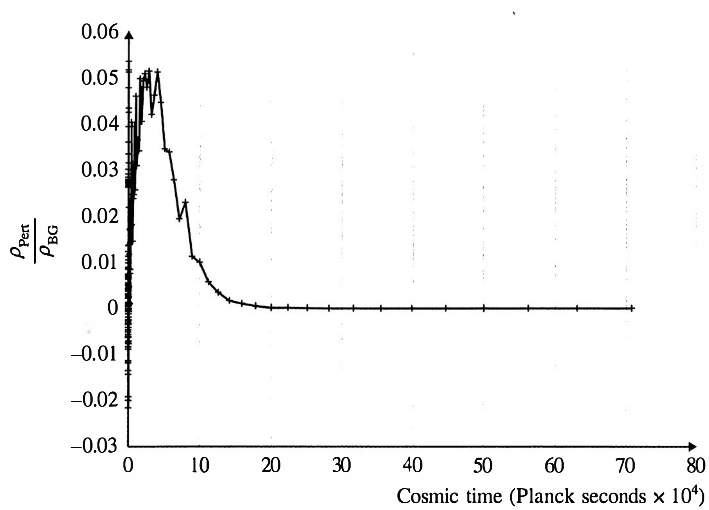


Fig. 39.5 Ratio of the energy density of scalar perturbation to the background energy density as a function of cosmic time. The initial conditions were chosen as $\phi_B = 1.23$ for the background, and the *obvious* fourth-order adiabatic vacuum at the bounce for perturbations. Slow-roll inflation starts about 3×10^5 Planck seconds after the bounce. During the entire evolution the ratio remains small. This example constitutes a self-consistent extension of the evolution of cosmic inhomogeneities to include the Planck era

of inflation differs significantly from the vacuum for modes in the interesting window and, as analyzed in detail in [39.97–101], those deviations have an important effect on the predictions of inflation for the spectrum of cosmic inhomogeneities, specially regarding non-Gaussianity. There exist concrete proposals for observables in the CMB [39.102, 104] and in the distribution of galaxies [39.102, 103] that should be sensitive to the effects of the created particles.

A quantity of direct observational interest is the power spectrum of tensor and scalar perturbations, $P_{\mathcal{T}}(k)$ and $P_{\mathcal{R}}(k)$ (see Chap. 30 for definitions), which are directly related to CMB observables. Figure 39.4 shows the relation between the LQC power spectrum computed with the evolved state and the standard inflationary power spectrum that assumes the BD-vacuum, for scalar perturbations. The conclusions are similar to the ones obtained from Fig. 39.3, namely for $\phi_B \gtrsim 1.2$ the power spectrum of observable modes is indistinguishable from the standard inflationary predictions. For smaller values of ϕ_B deviations become sizable for modes of observational interest. For instance, for $\phi_B = 1.15$ we have $k_{min} \approx 1$ and deviations from standard prediction will appear for modes with $\ell \lesssim 30$ in the WMAP angular decomposition. These deviations are inside current uncertainties. However, the fact that the state for perturbations differs from the BD-vacuum opens a window to observe those effects.

The analogous plot for tensor modes has the same form as Fig. 39.4, and the conclusions are also the

same [39.86]. In particular, there are no important corrections for the tensor-to-scalar ratio, although the inflationary consistency relation, which relates the tensor-to-scalar-ratio and the tensor spectral index, is modified [39.86].

Self-Consistency

The last step is to check whether there exist a big enough set of physical states ψ on the Hilbert space for which the truncation underlying our quantum theory, the test field approximation, holds during the entire evolution. This is an intricate question because:

i) It requires a detailed analytical control of the necessary regularization on states and composite operators on our Hilbert space.
ii) Numerical implementation of those techniques are necessary to check self-consistency *at any time during the evolution*, dealing with the subtleties of having numerical control on the subtraction of quantities that tend rapidly to infinity, during a period that covers around 11 orders of magnitude in energy density.

Section 39.3.2 summarized the necessary tools to check self-consistency and pointed out that a sufficient condition is that the energy density in perturbations $\langle\hat{\rho}\rangle$ be negligible compared to the background $\langle\hat{\rho}_0\rangle$ *at any time* during the evolution. Figure 39.5 shows the result of the numerical evolution of the energy density for scalar perturbations (analogous results hold for tensor perturbations). The plot shows the ratio $\langle\hat{\rho}_Q\rangle/\langle\hat{\rho}_0\rangle$ for a background corresponding to $\phi = 1.23$ and the *obvious* fourth adiabatic order vacuum specified at the bounce. This ratio remains small for the entire evolution, including the Planck regime. The initial condition $\phi = 1.23$ corresponds to $k_{\min} \approx 30$, therefore the number of excitations over the BD state on observable modes is negligible (Fig. 39.3) for this background. Additionally, there exist an analytical argument [39.86] ensuring that, given a state for perturbations for which back-reaction is negligible, there exist a well-defined neighborhood of that state with the same property. Each of those provide a state at the beginning of slow-roll indistinguishable from the BD vacuum. They provide therefore, viable extensions of the standard inflationary scenario that includes Planck scale physics [39.86, 87].

For the range $\phi_B < 1.2$ there are only upper bounds for $\langle\hat{\rho}_Q\rangle$ which are far from being optimal. At the present time there are no explicit computations for which the test field approximation is satisfied for ϕ_B in that window, and additional work is required to establish the self-consistency of our truncation scheme.

39.5 Conclusions

One of the most pressing questions a quantum theory of gravity has pertains to both theoretical and observational issues in cosmology. In the theoretical front the standard model is based on General Relativity that possesses an initial singularity, a signal that the theory breaks down at some point. On the observational front, the CMB spectrum poses very stringent conditions for any theory of the early universe. One of such scenarios is given by the inflationary paradigm, that explains very successfully the detailed structure of the inhomogeneities seen in the CMB as an imprint of quantum fluctuations of certain fields just before the inflationary phase. Can one have a formalism that provides a satisfactory, nonsingular description both at the Planck scale and at the onset of inflation? Interestingly, loop quantum cosmology allows one to answer both questions in the affirmative.

As we have described in this chapter, when one considers the homogeneous degrees of freedom, the so-called *background geometry*, the formalism provides precise singularity resolution, replacing the classical big bang with a big bounce. The dynamics of semiclassical states is very well described by an effective theory that captures the leading quantum gravity effects and allows one to describe the spacetime geometry in terms of an effective background metric.

The inflationary scenario is very powerful to explain in great detail many features of the observed CMB spectrum. It is however, incomplete in various directions. In particular, it is based on General Relativity where the spacetimes under consideration are past incomplete, that is, singular. As we have described in detail, one can indeed extend the scenario back in time to the Planck scale. For that one needs two new ingredients. The first one is a formalism that allows one to treat quantum perturbations of the spacetime metric propagating not on a classical spacetime, but rather on a *quantum* spacetime. The second ingredient in-

volves consistency conditions that ensure us that one can *evolve* the quantum perturbations back to the Planck scale without violating the approximations that yield validity to the formalism. As we have seen one can indeed consistently consider the extension of the inflationary scenario.

Perhaps the most pressing question is whether this extension to the quantum bounce provides a window for Planck scale physics to be observed in the CMB. As we have described, the sector of the parameter space that has been explored provides predictions that are fully consistent with the standard inflationary scenario, under current observations. Further explorations are needed to decide whether the scenario provided by LQC is both consistent in the full parameter space and provides us with distinct testable predictions.

References

39.1 A. Ashtekar, J. Lewandowski: Background independent quantum gravity: A status report, Class. Quantum Gravity **21**, R53–R152 (2004)

39.2 T. Thiemann: *Introduction to Modern Canonical Quantum General Relativity* (Cambridge Univ. Press, Cambridge 2007)

39.3 A. Ashtekar, M. Bojowald, J. Lewandowski: Mathematical structure of loop quantum cosmology, Adv. Theor. Math. Phys. **7**, 233–268 (2003)

39.4 A. Ashtekar, W. Kaminski, J. Lewandowski: Quantum field theory on a cosmological, quantum space-time, Phys. Rev. D **79**, 064030 (2009)

39.5 I. Agullo, A. Ashtekar, W. Nelson: An extension of the quantum theory of cosmological perturbations to the Planck era, Phys. Rev. D **87**, 043507 (2013)

39.6 A. Borde, A. Guth, A. Vilenkin: Inflationary spacetimes are not past-complete, Phys. Rev. Lett. **90**, 151301 (2003)

39.7 A. Ashtekar, M. Campiglia, A. Henderson: Casting loop quantum cosmology in the spin foam paradigm, Class. Quantum Gravity **27**, 135020 (2010)

39.8 A. Ashtekar, M. Campiglia, A. Henderson: Path integrals and the WKB approximation in loop quantum cosmology, Phys. Rev. D **82**, 124043 (2010)

39.9 C. Rovelli, F. Vidotto: On the spinfoam expansion in cosmology, Class. Quantum Gravity **27**, 145005 (2010)

39.10 E. Bianchi, C. Rovelli, F. Vidotto: Towards spinfoam cosmology, Phys. Rev. D **82**, 084035 (2010)

39.11 M. Martin-Benito, L.J. Garay, G.A. Mena Marugan: Hybrid quantum Gowdy cosmology: Combining loop and Fock quantizations, Phys. Rev. D **78**, 083516 (2008)

39.12 L.J. Garay, M. Martin-Benito, G.A. Mena Marugan: Inhomogeneous loop quantum cosmology: Hybrid quantization of the Gowdy model, Phys. Rev. D **82**, 044048 (2010)

39.13 D. Brizuela, G.A. Mena Marugan, T. Pawlowski: Big bounce and inhomogeneities, Class. Quantum Gravity **27**, 052001 (2010)

39.14 M. Martin-Benito, G.A. Mena Marugan, E. Wilson-Ewing: Hybrid quantization: From Bianchi I to the Gowdy model, Phys. Rev. D **82**, 084012 (2010)

39.15 M. Martin-Benito, D. Martin-de Blas, G.A. Mena Marugan: Matter in inhomogeneous loop quantum cosmology: the Gowdy T^3 model, Phys. Rev. D **83**, 084050 (2011)

39.16 D. Brizuela, G.A. Mena Marugan, T. Pawlowski: Effective dynamics of the hybrid quantization of the Gowdy T^3 universe, Phys. Rev. D **84**, 124017 (2011)

39.17 D. Brizuela, D. Cartin, G. Khanna: Numerical techniques in loop quantum cosmology, SIGMA **8**, 001 (2012)

39.18 M. Bojowald, G.M. Hossain: Loop quantum gravity corrections to gravitational wave dispersion, Phys. Rev. D **77**, 023508 (2008)

39.19 W. Nelson, M. Sakellariadou: Lattice refining loop quantum cosmology and inflation, Phys. Rev. D **76**, 044015 (2007)

39.20 J. Grain, A. Barrau: Cosmological footprints of loop quantum gravity, Phys. Rev. Lett. **102**, 081301 (2009)

39.21 J. Grain, T. Cailleteau, A. Barrau, A. Gorecki: Fully loop-quantum-cosmology-corrected propagation of gravitational waves during slow-roll inflation, Phys. Rev. D **81**, 024040 (2010)

39.22 J. Mielczarek, T. Cailleteau, J. Grain, A. Barrau: Inflation in loop quantum cosmology: Dynamics and spectrum of gravitational waves, Phys. Rev. D **81**, 104049 (2010)

39.23 J. Grain, A. Barrau, T. Cailleteau, J. Mielczarek: Observing the big bounce with tensor modes in the cosmic microwave background: Phenomenology and fundamental LQC parameters, Phys. Rev. D **82**, 123520 (2010)

39.24 M. Bojowald, G. Calcagni, S. Tsujikawa: Observational test of inflation in loop quantum cosmology, J. Cosmol. Astropart. Phys. **1111**, 046 (2011)

39.25 T. Cailleteau, J. Mielczarek, A. Barrau, J. Grain: Anomaly-free scalar perturbations with holonomy corrections in loop quantum cosmology, Class. Quantum Gravity **29**, 095010 (2012)

39.26 M. Fernandez-Mendez, G.A. Mena Marugan, J. Olmedo: Hybrid quantization of an inflationary universe, Phys. Rev. D. **86**, 024003 (2012)

39.27 E. Wilson-Ewing: Lattice loop quantum cosmology: Scalar perturbations, Class. Quantum Gravity **29**, 215013 (2012)

39.28 E. Wilson-Ewing: The matter bounce scenario in loop quantum cosmology, arXiv:1211.6269 (2013)

39.29 A. Ashtekar, P. Singh: Loop quantum cosmology: A status report, Class. Quantum Gravity **28**, 213001 (2011)

39.30 K. Banerjee, G. Calcagni, M. Martin-Benito: Introduction to loop quantum cosmology, SIGMA **8**, 016 (2012)

39.31 P. Singh: Numerical loop quantum cosmology: an overview, Class. Quantum Gravity **29**, 244002 (2012)

39.32 G. Calcagni: Observational effects from quantum cosmology, Ann. Phys. **525**, 323 (2013)

39.33 A. Ashtekar, T. Pawlowski, P. Singh: Quantum nature of the big bang: Improved dynamics, Phys. Rev. D **74**, 084003 (2006)

39.34 A. Ashtekar, A. Corichi, P. Singh: Robustness of predictions of loop quantum cosmology, Phys. Rev. D **77**, 024046 (2008)

39.35 A. Ashtekar, M. Campiglia: On the Uniqueness of Kinematics of Loop Quantum Cosmology, Class. Quantum Gravity **29**, 242001 (2012)

39.36 J. Lewandowski, A. Okolow, H. Sahlmann, T. Thiemann: Uniqueness of diffeomorphism invariant states on holonomy flux algebras, Commum. Math. Phys. **267**, 703–733 (2006)

39.37 C. Fleischhack: Representations of the Weyl algebra in quantum geometry, Commun. Math. Phys. **285**, 67–140 (2009)

39.38 A. Ashtekar, S. Fairhurst, J. Willis: Quantum gravity, shadow states, and quantum mechanics, Class. Quantum Grav. **20**, 1031–1062 (2003)

39.39 A. Ashtekar, T. Pawlowski, P. Singh: Quantum nature of the big bang: An analytical and numerical investigation, Phys. Rev. D **73**, 124038 (2006)

39.40 A. Corichi, P. Singh: Quantum bounce and cosmic recall, Phys. Rev. Lett. **100**, 209002 (2008)

39.41 W. Kaminski, T. Pawlowski: Cosmic recall and the scattering picture of loop quantum cosmology, Phys. Rev. D **81**, 084027 (2010)

39.42 A. Corichi, E. Montoya: On the semiclassical limit of loop quantum cosmology, Int. J. Mod. Phys. D **21**, 1250076 (2012)

39.43 A. Corichi, E. Montoya: Coherent semiclassical states for loop quantum cosmology, Phys. Rev. D **84**, 044021 (2011)

39.44 A. Ashtekar, T. Pawlowski, P. Singh, K. Vandersloot: Loop quantum cosmology of $k = 1$ FRW models, Phys. Rev. D **75**, 0240035 (2006)

39.45 L. Szulc, W. Kaminski, J. Lewandowski: Closed FRW model in loop quantum cosmology, Class. Quantum Gravity **24**, 2621 (2007)

39.46 A. Corichi, A. Karami: Loop quantum cosmology of $k = 1$ FRW: A tale of two bounces, Phys. Rev. D **84**, 044003 (2011)

39.47 K. Vandersloot: Loop quantum cosmology and the $k = -1$ RW model, Phys. Rev. D **75**, 023523 (2007)

39.48 L. Szulc: Open FRW model in Loop Quantum Cosmology, Class. Quantum Gravity **24**, 6191 (2007)

39.49 E. Bentivegna, T. Pawlowski: Anti-deSitter universe dynamics in LQC, Phys. Rev. D **77**, 124025 (2008)

39.50 A. Ashtekar, T. Pawlowski: Loop quantum cosmology with a positive cosmological constant, Phys. Rev. **85**, 064001 (2012)

39.51 W. Kaminski, T. Pawlowski: The LQC evolution operator of FRW universe with positive cosmological constant, Phys. Rev. D **81**, 024014 (2010)

39.52 A. Ashtekar, E. Wilson-Ewing: Loop quantum cosmology of Bianchi type I models, Phys. Rev. D **79**, 083535 (2009)

39.53 M. Martin-Benito, G.A. Mena Marugan, T. Pawlowski: Loop quantization of vacuum Bianchi I cosmology, Phys. Rev. D **78**, 064008 (2008)

39.54 M. Martin-Benito, G.A. Mena Marugan, T. Pawlowski: Physical evolution in loop quantum cosmology: The example of vacuum Bianchi I, Phys. Rev. D **80**, 084038 (2009)

39.55 A. Ashtekar, E. Wilson-Ewing: Loop quantum cosmology of Bianchi type II models, Phys. Rev. D **80**, 123532 (2009)

39.56 E. Wilson-Ewing: Loop quantum cosmology of Bianchi type IX models, Phys. Rev. D **82**, 043508 (2010)

39.57 A. Corichi, A. Karami: Loop quantum cosmology of Bianchi IX: Inverse triad corrections (unpublished)

39.58 A. Ashtekar, T.A. Schilling: Geometrical formulation of quantum mechanics. In: *On Einstein's Path: Essays in Honor of Engelbert Schücking*, ed. by A. Harvey (Springer, New York 1999) pp. 23–65

39.59 V. Taveras: LQC corrections to the Friedmann equations for a universe with a free scalar field, Phys. Rev. D **78**, 064072 (2008)

39.60 A. Corichi, T. Vukasinac: Effective constrained polymeric theories and their continuum limit, Phys. Rev. D **86**, 064019 (2012)

39.61 P. Singh: Are loop quantum cosmologies never singular?, Class. Quantum Gravity **26**, 125005 (2009)

39.62 A. Corichi, P. Singh: A geometric perspective on singularity resolution and uniqueness in loop quantum cosmology, Phys. Rev. D **80**, 044024 (2009)

39.63 P. Singh, K. Vandersloot, G.V. Vereshchagin: Nonsingular bouncing universes in loop quantum cosmology, Phys. Rev. D **74**, 043510 (2006)

39.64 A. Corichi, A. Karami: On the measure problem in slow roll inflation and loop quantum cosmology, Phys. Rev. D **83**, 104006 (2011)

39.65 A. Corichi, P. Singh: Is loop quantization in cosmology unique?, Phys. Rev. D **78**, 024034 (2008)

39.66 P. Singh, F. Vidotto: Exotic singularities and spatially curved loop quantum cosmology, Phys. Rev. D **83**, 064027 (2011)

39.67 V.A. Belinskii, I.M. Khalatnikov, E.M. Lifshitz: Oscillatory approach to a singular point in the relativistic cosmology, Adv. Phys. **31**, 525–573 (1970)

39.68 A. Ashtekar, A. Henderson, D. Sloan: Hamiltonian formulation of General Relativity and the Belinksii, Khalatnikov, Lifshitz conjecture, Class. Quantum Gravity **26**, 052001 (2009)
39.69 A. Ashtekar, A. Henderson, D. Sloan: A Hamiltonian formulation of the BKL conjecture, Phys. Rev. D **83**, 084024 (2011)
39.70 A. Corichi, A. Karami, E. Montoya: Loop quantum cosmology: Anisotropy and singularity resolution, arxiv:1210.7248 (2012)
39.71 A. Corichi, E. Montoya: Effective dynamics in Bianchi type II loop quantum cosmology, Phys. Rev. D **85**, 104052 (2012)
39.72 B. Gupt, P. Singh: Contrasting features of anisotropic loop quantum cosmologies: The role of spatial curvature, Phys. Rev. D **85**, 044011 (2012)
39.73 P. Singh: Curvature invariants, geodesics and the strength of singularities in Bianchi-I loop quantum cosmology, Phys. Rev. D **85**, 104011 (2012)
39.74 B. Gupt, P. Singh: Quantum gravitational Kasner transitions in Bianchi-I spacetime, Phys. Rev. D **86**, 024034 (2012)
39.75 J. Puchta: *Quantum fluctuations in quantum spacetime*, M.Sc. Thesis (Univ. Warsaw, Warsaw 2009)
39.76 A. Dapor, J. Lewandowski: Emergent isotropy-breaking in quantum cosmology, Phys. Rev. D **87**, 063512 (2012)
39.77 A. Dapor, J. Lewandowski, Y. Tavakoli: Lorentz symmetry in QFT on quantum Bianchi I spacetime, Phys. Rev. D **86**, 064013 (2012)
39.78 V.F. Mukhanov, H.A. Feldman, R.H. Brandenberger: Theory of cosmological perturbations, Phys. Rep. **215**(5/6), 203 (1992)
39.79 D. Langlois: Hamiltonian formalism and gauge invariance for linear perturbations in inflation, Class. Quantum Gravity **11**, 389–407 (1994)
39.80 B. Dittrich, J. Tambornino: Gauge invariant perturbations around symmetry reduced sectors of general relativity, Class. Quantum Gravity **24**, 4543–4585 (2007)
39.81 M. Bojowald, H.H. Hernandez, M. Kagan, P. Singh, A. Skirzewski: Hamiltonian cosmological perturbation theory with loop quantum gravity corrections, Phys. Rev. D **74**, 123512 (2006)
39.82 M. Bojowald, H.H. Hernandez, M. Kagan, P. Singh, A. Skirzewski: Formation and evolution of structure in loop cosmology, Phys. Rev. Lett. **98**, 031301 (2007)
39.83 K. Giesel, S. Hofmann, T. Thiemann, O. Winkler: Manifestly gauge-invariant general relativistic perturbation theory: I. Foundations, Class. Quantum Gravity **27**, 055005 (2010)
39.84 K. Giesel, S. Hofmann, T. Thiemann, O. Winkler: Manifestly gauge-invariant general relativistic perturbation theory: II. FRW background and first order, Class. Quantum Gravity **27**, 055006 (2010)
39.85 L. Bethke, J. Magueijo: Inflationary tensor fluctuations, as viewed by Ashtekar variables and their imaginary friends, Phys. Rev. D **84**, 024014 (2011)
39.86 I. Agullo, A. Ashtekar, W. Nelson: The pre-inflationary dynamics of loop quantum cosmology: Confronting quantum gravity with observations, Class. Quantum Gravity **30**, 085014 (2013)
39.87 I. Agullo, A. Ashtekar, W. Nelson: A quantum gravity extension of the inflationary scenario, Phys. Rev. Lett. **109**, 251301 (2012)
39.88 E. Komatsu, K.M. Smith, J. Dunkley, C.L. Bennett, B. Gold, G. Hinshaw, N. Jarosik, D. Larson, M.R. Nolta, L. Page, D.N. Spergel, M. Halpern, R.S. Hill, A. Kogut, M. Limon, S.S. Meyer, N. Odegard, G.S. Tucker, J.L. Weiland, E. Wollack, E.L. Wright: Seven-year Wilkinson microwave anisotropy probe (WMAP) observations: Cosmological interpreation, Astrophys. J. Suppl. Ser. **192**, 18 (2011)
39.89 A. Ashtekar, D. Sloan: Probability of inflation in loop quantum cosmology, Gen. Relativ. Gravit. **43**, 3619–3656 (2011)
39.90 G.W. Gibbons, S.W. Hawking, J. Stewart: A natural measure on the set of all universes, Nucl. Phys. B **281**, 736 (1987)
39.91 S.W. Hawking, D.N. Page: How probable is inflation?, Nucl. Phys. B **298**, 789 (1988)
39.92 J.S. Schiffrin, R.M. Wald: Measure and probability in cosmology, Phys. Rev. D **86**, 023521 (2012)
39.93 G.W. Gibbons, N. Turok: The measure problem in cosmology, Phys. Rev. D **77**, 063516 (2008)
39.94 L. Parker: The Creation of Particles in an Expanding Universe (Harvard Univ., Cambridge 1966)
39.95 L. Parker: Particle creation in expanding universes, Phys. Rev. Lett. **21**, 562 (1968)
39.96 L. Parker: Quantized fields and particle creation in expanding universes 1, Phys. Rev. **183**, 1057 (1969)
39.97 R. Holman, A. Tolley: Enhanced Non-Gaussianity from excited states, J. Cosmol. Astropart. Phys. **0805**, 001 (2008)
39.98 I. Agullo, L. Parker: Non-Gaussianities and the stimulated creation of quanta in the inflationary universe, Phys. Rev. D **83**, 063526 (2011)
39.99 I. Agullo, L. Parker: Stimulated creation of quanta during inflation and the observable universe, Gen. Relativ. Gravit. **43**, 2541–2545 (2011)
39.100 J. Ganc: Calculating the local-type fNL for slow-roll inflation with a non-vacuum initial state, Phys. Rev. D **84**, 063514 (2011)
39.101 I. Agullo, J. Navarro-Salas, L. Parker: Enhanced local-type inflationary trispectrum from a non-vacuum initial state, J. Cosmol. Astropart. Phys. **1205**, 019 (2012)
39.102 J. Ganc, E. Komatsu: Scale dependent bias of galaxies and μ-type distortion of the cosmic microwave background spectrum from asingle field inflation with a modified initial state, Phys. Rev. D **86**, 023518 (2012)

39.103 I. Agullo, S. Shandera: Large non-Gaussian halo bias from single field inflation, J. Cosmol. Astropart. Phys. **1209**, 007 (2012)
39.104 F. Schmidt, L. Hui: CMB power asymmetry from Gaussian modulation, Phys. Rev. Lett. **110**, 011301 (2013)
39.105 R. Penrose: Singularities and time-asymmetry. In: *General Relativity: An Einstein Centenary Survey*, ed. by S.W. Hawking, W. Israel (Cambridge Univ. Press, Cambridge 1979) pp. 581–638

Acknowledgements

A.3 Relativity Today

by Nick M. J. Woodhouse

Part of this chapter was taken from: N.M.J. Woodhouse, *Special Relativity*, Springer Undergraduate Mathematics Series (Springer, London, 2003)

A.4 Acceleration and Gravity: Einstein's Principle of Equivalence

by Lewis Ryder

It is my pleasure to acknowledge helpful conversations and correspondence on this subject with Domenico Giulini, Robert Low, Steve Lyle, Bahram Mashhoon, Bijan Sheikholeslami-Sabzevari, and Robin Tucker, and to thank Volker Perlick for his kind invitation to the Bad Honnef seminar on Problems and Developments of Classical Electrodynamics, where these conversations took place.

B.10 The Nature and Origin of Time-Asymmetric Spacetime Structures

by H. Dieter Zeh

I wish to thank Claus Kiefer for his comments on an early draft of this manuscript.

B.11 Teleparallelism: A New Insight into Gravity

by José G. Pereira

The author would like to thank R. Aldrovandi for his long-standing collaboration in the development of the ideas presented in this monograph. He would like to thank also FAPESP, CAPES and CNPq for partial financial support.

B.12 Gravity and the Spacetime: An Emergent Perspective

by Thanu Padmanabhan

The author would like to thank Sunu Engineer for several discussions. The research is partially supported by J.C. Bose research grant of DST, India.

B.13 Spacetime and the Passage of Time

by George Ellis, Goswami Rituparno

We thank C. Clarkson, R. Tavakol, and T. Clifton for helpful comments.

C.16 The Initial Value Problem in General Relativity

by James Isenberg

This work was partially supported by NSF grant PHY-0968612 at the University of Oregon.

C.19 Conserved Charges in Asymptotically (Locally) AdS Spacetimes

by Donald Marolf, William Kelly, Sebastian Fischetti

This work was supported in part by the National Science Foundation under Grant Nos PHY11-25915 and PHY08-55415, and by funds from the University of California. DM also thanks the University of Colorado, Boulder, for its hospitality during this work.

D.25 Quasi-local Black Hole Horizons

by Badri Krishnan

I am grateful to Abhay Ashtekar for valuable discussions and suggestions. I also thank Ingemar Bengtsson and Jose Senovilla for valuable comments.

D.26 Gravitational Astronomy

by B. Suryanarayana Sathyaprakash

In writing this review I benefitted from discussions with Bernard Schutz. I would like to thank Professor V. Petkov for inviting me to write this review and for his patience and encouragement to finish it.

D.27 Probing Dynamical Spacetimes with Gravitational Waves

by Chris Van Den Broeck

The author is supported by the research programme of the Foundation for Fundamental Research on Matter (FOM), which is partially supported by the Netherlands Organization for Scientific Research (NWO). It is a pleasure to thank M. Agathos, K.G. Arun, J.F.J. van den Brand, N. Cornish, W. Del Pozzo, K. Grover, M. Hendry, I.S. Heng, B.R. Iyer, T.G.F. Li, I. Mandel, C. Messenger, C.K. Mishra, A. Pai, M. Pitkin, B.S. Sathyaprakash, B.F. Schutz, P.S. Shawhan, T. Sidery, R. Sturani, M. Tompitak, M. Vallisneri, A. Vecchio, J. Veitch, S. Vitale, and N. Yunes, for fruitful discussions. I would like to acknowledge the LIGO

Data Grid clusters, without which some of the simulations described here could not have been performed. Specifically, these include the computing resources supported by National Science Foundation awards PHY-0923409 and PHY-0600953 to UW-Milwaukee. Also, I thank the Albert Einstein Institute in Hannover, supported by the Max-Planck-Gesellschaft, for use of the Atlas high-performance computing cluster.

E.28 Einstein's Equations, Cosmology, and Astrophysics

by Paul S. Wesson

Thanks go to the students who in the past asked good questions and to the colleagues who shared their research, notably on higher-dimensional relativity (http://www.5dstm.org/).

E.31 Exact Approach to Inflationary Universe Models

by Sergio del Campo

This work was supported by the Comision Nacional de Ciencias y Tecnologia through FONDECYT Grant No. 1 110 230 and also was partially supported by PUCV Grant No. 123 710.

F.34 Quantum Gravity via Causal Dynamical Triangulations

by Jan Ambjørn, Andrzej Görlich, Jerzy Jurkiewicz, Renate Loll

JA and AG thank the Danish Research Council for financial support via the grant *Quantum gravity and the role of black holes*, and the EU for support through the ERC Advanced Grant 291092, *Exploring the Quantum Universe* (EQU). JJ acknowledges a partial support of the International PhD Projects Program of the Foundation for Polish Science within the European Regional Development Fund of the European Union, agreement no. MPD/2009/6 as well as support from grant DEC-2012/06/A/ST2/00389 from the National Science Centre Poland. RL acknowledges support through several Projectruimte grants by the Dutch Foundation for Fundamental Research on Matter (FOM).

F.38 Spin Foams

by Jonathan S. Engle

The author thanks his wife, Sabine Engle, for careful assistance with the figures in this chapter, the editors and an anonymous referee for assistance in improving the chapter, and Christopher Beetle for pointing out reference [39.40]. This work was supported in part by the National Science Foundation through grant PHY-1237510 and by the National Aeronautics and Space Administration through the University of Central Florida's NASA–Florida Space Grant Consortium.

F.39 Loop Quantum Cosmology

by Ivan Agullo, Alejandro Corichi

We would like to thank A. Ashtekar, P. Singh, and W. Nelson for discussions and collaboration. I.A. thanks the Marie Curie program of the EU for funding. This work was partly funded by DGAPA-UNAM IN103610, CONACyT CB0177840, and NSF PHY0854743 grants and by the Eberly Research Funds of Penn State.

About the Authors



Ivan Agullo

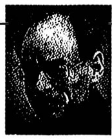
University of Cambridge
Department of Applied Mathematics and Theoretical Physics
Cambridge, UK
Louisiana State University
Department of Physics & Astronomy
Baton Rouge, USA
i.agullorodenas@damtp.cam.ac.uk

Chapter F.39

Ivan Agullo studied Physics at the University of Valencia, Spain, and obtained a PhD in theoretical Physics in 2009. Currently, Dr. Agullo is a Marie Curie Postdoctoral Fellow at the Department of Applied Mathematics and Theoretical Physics, University of Cambridge. He joined Louisiana State University in August 2013 as an Assistant Professor of Physics. Dr. Agullo's research has focused on the interplay between gravitation and the quantum theory, with contributions to the theory of quantized fields in curved space-time and loop quantum gravity, especially to the physics of black holes and the early universe.

Jan Ambjørn

Copenhagen University
The Niels Bohr Institute
Copenhagen, Denmark
Radboud University Nijmegen
Institute for Mathematics, Astrophysics and Particle Physics (IMAPP)
Nijmegen, Netherlands
ambjorn@nbi.dk



Chapter F.34

Jan Ambjørn earned his PhD from the Niels Bohr Institute, Copenhagen in 1980, had postdoc positions at Caltech, Pasadena and Nordita, Stockholm. He was Professor at the Niels Bohr Institute from 1985 – 2003, at Utrecht University 2003-2010, and is Professor at Radboud University, Nijmegen since 2012. His research areas are in quantum field theory, random surfaces, and string theory as well as quantum gravity.

Neil Ashby

NIST
Time and Frequency Division
Boulder, USA
ashby@boulder.nist.gov



Chapter D.24

Professor Neil Ashby received a BA degree (Summa Cum Laude) in Physics from the University of Colorado, Boulder, in 1955, MSc and PhD degrees from Harvard University, Cambridge. In 1962, he joined the Faculty of the Department of Physics at the University of Colorado. He has been a Professor of Physics there since 1970 and was Department Chair from 1984–1988. He is an Associate at the Time and Frequency Division of NIST, working on relativistic effects on clocks, global time synchronization, and on quantum effects in atomic fountains. In 2005 he received the F.K. Richtmyer Award from the American Association of Physics Teachers.



Beverly K. Berger

Livermore, USA
beverlyberger@me.com

Chapter C.21

Dr. Beverly K. Berger retired from the US National Science Foundation where she was Program Director for Gravitational Physics from late 2001 until the end of 2011. She had spent 24 years as a faculty member at Oakland University (Michigan, USA), eventually becoming Professor and Department Chair. She received her undergraduate education in Physics at the University of Rochester and her PhD in Physics at the University of Maryland. Her research field is theoretical gravitational physics with recent emphasis on singularities and other properties of cosmological spacetimes. In 1995, she founded the American Physical Society's Topical Group in Gravitation (GGR) and was elected Vice Chair of GGR in 2012.



Orfeu Bertolami

Chapter D.22

Universidade do Porto
Faculdade de Ciências, Departamento de Física e Astronomia
Porto, Portugal
orfeu.bertolami@fc.up.pt

Orfeu Bertolami is Professor of Theoretical Physics at Departamento de Física e Astronomia of the Faculdade de Ciências of Universidade do Porto, Portugal. He received his PhD in Theoretical from the University of Oxford, UK in 1987, and M.Sc. degrees from the University of Cambridge, UK, and Instituto de Física Teórica in São Paulo, Brazil. His research interests include Cosmology, Classical and Quantum Gravity, Mathematical and Physical Foundations of Quantum Mechanics and Fundamental Physics in Space. He has been member of the Galileo Science Advisory Committee of the European Space Agency (ESA).

Robert T. Bluhm

Colby College
Department of Physics and Astronomy
Waterville, USA
rtbluhm@colby.edu



Chapter D.23

Robert Bluhm is the Sunrise Professor of Physics at Colby College in Waterville, Maine. He received his PhD in theoretical physics from The Rockefeller University in 1988, and was a postdoc at Indiana University. He has been at Colby College since 1990. Robert's main interests are in fundamental symmetries in particle physics and gravity. Much of his work has been on finding new and improved ways to test Lorentz symmetry.

Sergio del Campo

Pontificia Universidad Catolica de Valparaiso
Valparaíso, Chile
sdelcamp@ucv.cl



Chapter E.31

Sergio del Campo is a Professor of Physics at the Pontifical Catholic University of Valparaíso, Chile. His research covers the early universe, specifically inflationary universe models, the recent accelerated evolution of the universe, the physics of wormhole and black holes. He is also engaged in research on quantum cosmology, where the study of the tunneling wave function of the universe was his doctoral thesis at Tufts University.



Alejandro Corichi

Chapter F.39

National Autonomous University of Mexico (UNAM)
Centro de Ciencias Matematicas, Quantum Gravity Group, UNAM Campus Morelia
Morelia, Michoacan, Mexico

Alejandro Corichi is a Professor of Gravitational Physics at the Center for Mathematical Sciences, UNAM-Morelia, Mexico. He graduated in Physics from the Universidad Nacional Autónoma de México (UNAM) in 1991, obtained a PhD from The Pennsylvania State University in 1997. He has held visiting positions at the University of Mississippi and Penn State. His area of research is the strong gravity regime in classical and quantum gravity, in particular black holes and the very early universe. He is a member of the Mexican Academy of Sciences and a Fellow of the International Society on General Relativity and Gravitation.



Sergio Dain

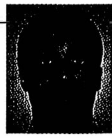
Chapter C.18

Universidad Nacional de Córdoba
Facultad de Matemática, Astronomía y Física
Ciudad Universitaria, Córdoba, Argentina
dain@famaf.unc.edu.ar

Sergio Dain studied Physics at the National University of Córdoba. From 1994 to 1999 he worked on his thesis at the Córdoba University and at the Max Planck Institute for Gravitational Physics, Potsdam, Germany. In 1999 he obtained his PhD from the National University of Córdoba. After six years as a post-doc at the Max Planck Institute for Gravitational Physics, he is now Professor at the National University of Córdoba and Member of the Argentina National Scientific and Technological Research Council (CONICET). Sergio Dain's current research is mainly concerned with geometrical inequalities for black holes.

Diako Darian

Oslo, Norway
diako.darian@gmail.com



Chapter E.29

Diako Darian is an astrophysicist from Norway. He graduated from University of Oslo specializing in cosmology and universe models in spring 2010. In his master thesis, he thoroughly investigated the role of viscosity in the evolution of the Universe under the assistance of his supervisor, Professor Øyvind Grøn. After his thesis he has continued his work on the impact of the viscosity on the Universe.

Dennis Dieks

Utrecht University
History and Foundations of Science
Utrecht, Netherlands
d.dieks@uu.nl



Chapter

Dennis Dieks studied theoretical physics in Amsterdam and obtained his PhD in the foundations of physics at Utrecht University. He is a Professor of Philosophy and Foundations of the Natural Sciences in Utrecht, a member of the Royal Netherlands Academy of Sciences, editor of Studies in the History and Philosophy of Modern Physics and associate editor of Foundations of Physics.



George F.R. Ellis

University of Cape Town
Department of Mathematics
Rondebosch, Cape Town, South Africa
george.ellis@uct.ac.za

Chapter B.13

George Ellis is Professor Emeritus of Applied Mathematics, University of Cape Town, South Africa. His professional research work is in relativity theory, cosmology, and complexity studies. After a BSc (Hons) at the University of Cape Town he did a PhD in Applied Mathematics and Theoretical Physics at Cambridge University. He has been Professor of Applied Mathematics at the University of Cape Town and Visiting Professor at the University of Alberta, Edmonton, University of Texas, Austin, Professor of Cosmic Physics, at the International School of Advanced Studies (SISSA), Trieste, Italy, and G C MacVittie Visiting Professor of Astronomy, and Queen Mary College, London University.



Jonathan S. Engle

Florida Atlantic University
Department of Physics
Boca Raton, USA
jonathan.engle@fau.edu

Chapter F.38

Jonathan Steven Engle is currently an Assistant Professor in the Spacetime Physics group at Florida Atlantic University. He obtained his PhD in Physics from the Pennsylvania State University and has held postdoctoral positions at the Centre de Physique Théorique in Marseille, the Albert-Einstein-Institute in Potsdam, Germany, and the Friedrich-Alexander-Universität in Erlangen, Germany. As of 2012 he also holds an affiliate professor position at the Friedrich-Alexander-Universität. His principal research interests center on general relativity, and in particular on loop quantum gravity and its applications to black holes and cosmology.

Rafael Ferraro

Instituto de Astronomía y Física del Espacio
Buenos Aires, Argentina
Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales,
Departamento de Física
Buenos Aires, Argentina
ferraro@iafe.uba.ar



Chapter A.1

Rafael Ferraro is a Professor at the Department of Physics, Universidad de Buenos Aires (UBA), and a research scientist at the Instituto de Astronomía y Física del Espacio (CONICET-UBA), Argentina. He is a graduate from UBA (1982), where he also received his PhD (1986). He is the author of a textbook on Special and General Relativity. In the last years, his research work has been focused on theories of modified gravity and nonlinear electrodynamics. He has also researched in areas such as quantum theory in curved space, quantum gravity, constrained Hamiltonian systems.

Sebastian Fischetti

University of California Santa Barbara
Department of Physics
Santa Barbara, USA
sfischet@physics.ucsb.edu



Chapter C.19

Sebastian Fischetti is a graduate student of Don Marolf in the Physics department at the University of California, Santa Barbara. He studies strongly interacting fields in curved spacetime and gravitational aspects of string theory.



Maurizio Gasperini

University of Bari
Department of Physics
Bari, Italy
gasperini@ba.infn.it

Chapter F.35

Maurizio Gasperini graduated in Physics from the University of Bologna (1975) and has served for many years as a researcher in theoretical physics at the University of Turin (1983–1998). He is currently Professor of Theoretical Physics at the University of Bari, member of the Committee of the Italian Ministry of Education (MIUR) for the *National Scientific Qualification*, member of the International Board of Experts for the *Evaluation of Research Quality* of MIUR, scientific expert of the French National research Agency (ANR) for the *Blanc Program* of scientific excellence. His main teaching and research activities concern relativity, cosmology, gravitational theory and unified theories of fundamental interactions.



Domenico Giulini

Leibniz Universität Hannover
Institut für Theoretische Physik
Hannover, Germany
giulini@itp.uni-hannover.de

Chapter C.17

Domenico Giulini studied Physics and Mathematics at Heidelberg University and moved to Cambridge (England) to do part III of the mathematical tripos and subsequently a PhD under the supervision of Gary Gibbons. He returned to Germany to take up a position in the group of Klaus Pohlmeyer at Freiburg University, where he received his habilitation in 1996. He then worked at Zurich University and the Albert Einstein Institute at Golm. He currently holds a professorship in theoretical physics at the Leibniz University Hannover (Germany),where he is a member of the Riemann Center for Geometry and Physics.

Andrzej Görlich

Copenhagen University
The Niels Bohr Institute
Copenhagen, Denmark
Jagellonian University
Marian Smoluchowski Institute of Physics,
Department of Theory of Complex Systems
Kraków, Poland
atg@th.if.uj.edu.pl



Chapter F.34

Andrzej Görlich received his PhD from the Jagiellonian University in 2010, was postdoc at the Jagiellonian University 2010-2011 and at the Niels Bohr Institute since 2011. His research areas are quantum gravity, random matrix theory, quantum computing.

Øyvind Grøn

Oslo and Akershus University College of Applied Sciences
Oslo, Norway
University of Oslo
Institute of Physics
Oslo, Norway
oyvind.gron@hioa.no



Chapters B.9, E.29

Øyvind Grøn is Professor of Physics at Oslo and Akershus University College of Applied Sciences and Professor II at the University of Oslo. Grøn has conducted research within the areas of general relativity, cosmology and classical electromagnetism. He has also studied properties of the electromagnetic field produced by accelerated electric charges, and the properties of conformally flat spacetimes. In several studies Grøn has focused on relativistic models of the universe. He and his co-workers have shown that it is possible to interpret observational data so that dark energy is not necessary.



Graham S. Hall

University of Aberdeen
Institute of Mathematics
Aberdeen, UK
g.hall@abdn.ac.uk

Chapter A.5

Graham Hall is Emeritus Professor of Mathematics in the Institute of Mathematics at The University of Aberdeen, Scotland. He was a student at the University of Newcastle Upon Tyne, England, and was the Earl Grey Memorial Fellow there from 1971–1973. His main areas of interest lie in differential geometry and classical mathematical relativity theory. His research has mainly concentrated on symmetries in general relativity theory, the classification of gravitational fields, holonomy groups, curvature structure and projective structure. He is a Fellow of The Royal Society of Edinburgh (1995) and a Fellow of The Royal Astronomical Society (2004).



James Isenberg

University of Oregon
Department of Mathematics
Eugene, USA
isenberg@uoregon.edu

Chapter C.16

James Isenberg is a Professor of Mathematics and a member of the Institute for Theoretical Science at the University of Oregon. He is a Fellow of the American Physical Society and holds degrees from Princeton University and the University of Maryland. His main areas of research are mathematical general relativity and geometric heat flows. He has been a leader in the study of solutions of the Einstein constraint equations, in the analysis of cosmological singularities in spacetime solutions of Einstein's equations, and in modeling neckpinch singularities in Ricci flow.

Pankaj S. Joshi

Tata Institute of Fundamental Research
Mumbai, India
psj@tifr.res.in



Chapter C.20

Pankaj S. Joshi works in gravitation and cosmology and is a Senior Professor at the Tata Institute of Fundamental Research, Mumbai. He did postdoctoral work at the University of Pittsburg and spent some time at the University of Cambridge. He worked on various basic issues related to gravitational collapse, cosmic censorship, and formation of black holes and naked singularities as collapse end states, which are issues at the foundation of black hole physics and modern applications in relativistic astrophysics. His Scientific American article on *Naked Singularities* in 2009, drew much attention and was translated in many international languages. He has received many honors and awards and is a Fellow of the Indian National Science Academy and the National Academy of Sciences, India.

Jerzy Jurkiewicz

Jagellonian University
Marian Smoluchowski Institute of Physics, Department of Theory of Complex Systems
Kraków, Poland
jurkiewicz@th.if.uj.edu.pl



Chapter F.34

PhD from Jagiellonian University 1975, postdoc at Utrecht University and Paris-Sud University, Professor at Jagiellonian University since 1996, visiting Professor at the Niels Bohr Institute 1990–1991, 1994–1996. His research areas: quantum field theory (lattice regularization), random matrix theory, quantum gravity, complex systems.



William Kelly

University of California Santa Barbara
Department of Physics
Santa Barbara, USA
wkelly@physics.ucsb.edu

Chapter C.19

William Kelly is a graduate student at the University of California, Santa Barbara where he studies gravitational aspects of AdS/CFT.



Claus Kiefer

University of Cologne
Institute for Theoretical Physics
Köln, Germany
kiefer@thp.uni-koeln.de

Chapter F.33

Professor of Theoretical Physics at the University of Cologne, Claus Kiefer does research on quantum gravity and the foundations of quantum theory as well as studying black holes and more general cosmological questions. He was educated at the University of Vienna and at the University of Heidelberg, where he completed his undergraduate work and went on to take a PhD in physics summa cum laude in 1988. After postdoctoral research at the University of Zurich, he joined the physics faculty of the University of Freiburg as a lecturer in 1993. He was named to his present position at the University of Cologne in 2001.

Badri Krishnan

Max Planck Institute for Gravitational Physics
Observational Relativity and Cosmology
Hannover, Germany
badri.krishnan@aei.mpg.de



Chapter D.25

Badri Krishnan obtained his MSc degree from the Indian Institute of Technology, Kanpur in 1997 and his PhD degree from the Pennsylvania State University in 2002. He is currently a senior staff scientist at the Max-Planck-Insitute for Gravitational Physics in Hannover, Germany. His research interests include searching for gravitational waves and in theoretical aspects of general relativity.

Renate Loll

Radboud University Nijmegen
Institute for Mathematics, Astrophysics and Particle Physics (IMAPP)
Nijmegen, Netherlands
r.loll@science.ru.nl



Chapter F.34

PhD Imperial College London, UK (1989); Habilitation, Potsdam, Germany (1998); Heisenberg Fellow (D); Vici Fellow (NL); Professor at Utrecht University (2005–2012); Professor at Radboud University Nijmegen (since 2012); Distinguished Research Chair, Perimeter Institute for Theoretical Physics (since 2009).



Stephen N. Lyle

Alzen, France
stephen.n.lyle@gmail.com

Chapter B.7

Stephen Lyle is a freelance science editor based in France. He studied pure mathematics and theoretical physics at Trinity College, Cambridge, UK, and noncommutative geometry and integration at the University of Paris VI, France. He has published two monographs on theoretical physics.



Donald Marolf

University of California Santa Barbara
Department of Physics
Santa Barbara, USA
marolf@physics.ucsb.edu

Chapter C.19

Don Marolf is a Professor of Physics at the University of California, Santa Barbara, where he studies the quantum physics of black holes, interacting fields in curved space, and gravitational aspects of string theory.

Thanu Padmanabhan

IUCAA
Pune University Campus
Pune, India
paddy@iucaa.ernet.in

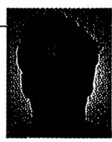


Chapter B.12

Professor Padmanabhan, Distinguished Professor and Dean at IUCAA, Pune, is renowned for his research contributions to the subjects of gravitation and cosmology. He provided a clear interpretation of gravity as an emergent phenomenon and showed that this paradigm extends to a wide class of gravitational theories including, but not limited to, Einstein's theory. He has authored nine advanced-level textbooks, acclaimed and used worldwide as standard references. He is currently the Chairman of the IUPAP Astrophysics Commission and was the President of the IAU Cosmology Commission and a Sackler Distinguished Astronomer of the Institute of Astronomy, Cambridge. In recognition of his achievements, the President of India awarded him the medal of honor, Padma Shri, in 2007.

Jorge Páramos

Universidade do Porto
Faculdade de Ciências, Departamento de Física e Astronomia
Porto, Portugal
jorge.paramos@fc.up.pt



Chapter D.22

Jorge Páramos is a researcher, focusing mainly on phenomenological implications of modifications of gravity and the design of scientific objectives for space missions.



José G. Pereira

Universidade Estadual Paulista
Instituto de Física Teórica
São Paulo, Brazil
jpereira@ift.unesp.br

Chapter B.11

José Geraldo Pereira graduated from Universidade de São Paulo in 1979, and got his PhD from the Instituto de Física Teórica in 1986. From 1987 to 1989 he was a (Brazilian funded) post-doctoral fellow at the University of Texas at Austin, USA. Currently, he is Associate Professor at the Instituto de Física Teórica, Universidade Estadual Paulista, in São Paulo, Brazil.



Vesselin Petkov

Institute for Foundational Studies
Hermann Minkowski
Montreal, Quebec, Canada
vpetkov@minkowskiinstitute.org

Chapter B.8

Vesselin Petkov received a doctorate in philosophy from the Institute for Philosophical Research of the Bulgarian Academy of Sciences, and a doctorate in physics from Concordia University in Montreal. He is one of the founding members and the current Director of the Institute for Foundational Studies Hermann Minkowski.

Goswami Rituparno

University of KwaZulu-Natal
School of Mathematics Statistics and
Computer Science, Westville Campus
Durban, South Africa
vitasta9@gmail.com



Chapter B.13

Rituparno Goswami is a Senior Lecturer in the School of Mathematics Statistics and Computer Science, University of KwaZulu-Natal, South Africa. His professional research work is in general relativity and cosmology. After his BSc (Hons) and MSc from Indian Institute of Technology (IIT) Kharagpur he did his PhD on theory of gravitation at Tata Institute of Fundamental Research (TIFR), India. He has been Postdoctoral Fellow at the University of Alberta, Edmonton and the University of Cape Town.

Carlo Rovelli

Aix-Marseille University
Centre de Physique Théorique de
Luminy
Marseille, France
rovelli@cpt.univ-mrs.fr



Chapter F.36

Carlo Rovelli got his PhD in Padova in 1986. He has worked in the Universities of Rome, Yale and Pittsburgh, and is currently Professor of Theoretical Physics in Marseille. He is a Honorary Professor of the Normal University of Beijing, member of the Institut Universitaire de France and of the Académie Internationale de Philosophie des Sciences. He has received the 1995 Xanthopoulos Prize for the foundation of Loop Quantum Gravity and his work on the nature of space and time.



Lewis Ryder

University of Kent
School of Physical Sciences, Ingram
Building
Canterbury, UK
l.h.ryder@kent.ac.uk

Chapter A.4

Lewis Ryder graduated from Oxford University and took his PhD at Edinburgh University in Theoretical Particle Physics. One year of this was spent at the Middle East Technical University in Ankara, studying under Feza Gürsey. Then he was appointed to Lectureship at Kent University in Canterbury where he has remained. He is the author of books *Elementary Particles and Symmetries* (1975, 1984); *Quantum Field Theory* (1985, 1996) and *Introduction to General Relativity* (2009).



Hanno Sahlmann

Friedrich-Alexander University
Erlangen-Nürnberg
Department of Physics
Erlangen, Germany
hanno.sahlmann@gravity.fau.de

Chapter F.37

Hanno Sahlmann obtained his PhD at the Max Planck Institute for Gravitational Physics in Potsdam (Germany) in 2002. He held postdoctoral positions in the USA, the Netherlands, and Germany, and led an independent research group of the Max-Planck-Gesellschaft in South Korea. Since 2012 he is a Professor at the University Erlangen-Nürnberg. His work is concerned with the interplay of geometry and quantum theory. This includes quantum field theory on curved space-times, general relativity, alternative theories of gravitation, and the quantization of gravitational field itself.

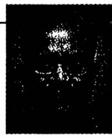
B. Suryanarayana Sathyaprakash

Cardiff University
School of Physics and Astronomy
Cardiff, UK
b.sathyaprakash@astro.cf.ac.uk



Chapter D.26

Sathyaprakash is a Professor at the School of Physics and Astronomy at Cardiff University, UK and Head of its Gravitational Physics Group. His research interests include classical field theory and its relevance to cosmology, formation and evolution of large-scale structure in the Universe, and statistical, morphological and geometrical properties of large-scale structure. He is involved in all the major interferometric gravitational wave projects including American LIGO, French-Italian Virgo and the British-German GEO600. He has also been involved in the European Design Study of a third-generation detector called the Einstein Telescope.



Tarun Souradeep

Inter-University Centre for Astronomy and Astrophysics (IUCAA)
Pune, India
tarun@iucaa.ernet.in

Chapter E.32

Souradeep graduated as an engineer from IIT Kanpur, India. After a short stint in automobile design, he decided to pursue a career in Gravitation and Cosmology at IUCAA, CITA and KSU. Since 2000, he has lead a group on Cosmic Microwave background at IUCAA. He is a core team member of the Planck CMB mission and also deeply involved in the LIGO-India project proposal.



Norbert Straumann

University of Zurich
Institute for Theoretical Physics
Zurich, Switzerland
norbert.straumann@gmail.com

Chapter

Norbert Straumann studied physics and mathematics at the ETH Zurich. In 1969, Straumann became professor at the University of Zurich, followed by visiting professorships at the University of Bern and Amsterdam. From 1997–2000 he worked on an advisory board of the Albert Einstein Institute in Potsdam. In 2001 he retired, and has since written a number of further works on the theory of gravitation, primarily addressing the classic fields of theoretical physics. His lecture notes, which have partly been published in book form, influenced a whole generation of students and researchers at the University of Zurich. In recognition of his achievements, in 2005 Norbert Straumann was awarded the title of Doctor Philosophiae Honoris Causa from the University of Bern.



Chris Van Den Broeck

National Institute for Subatomic Physics (Nikhef)
Department of Gravitational Physics
Amsterdam, Netherlands
vdbroeck@nikhef.nl

Chapter D.27

Chris Van Den Broeck is a Working Group Leader at the National Institute for Subatomic Physics (Nikhef) in Amsterdam, The Netherlands, and chair of the Virgo Collaboration's compact binary coalescence data analysis group. His research focuses on developing algorithms to extract science from gravitational wave observations that will soon be made by the Advanced Virgo and Advanced LIGO observatories.

Scott Walter

University of Lorraine
LHSP-Archives Henri-Poincaré (CNRS, UMR 7117)
Nancy, France



Chapter A.2

Scott Walter directs the graduate program in history and philosophy of science at the University of Lorraine, edits Henri Poincaré's scientific correspondence, and pursues research in the history of modern mathematical sciences.

David Wands

University of Portsmouth
Institute of Cosmology and Gravitation
Portsmouth, UK
david.wands@port.ac.uk



Chapter E.30

David Wands is Director of the Institute of Cosmology and Gravitation at the University of Portsmouth. His research investigates the very early Universe and the origin of structure. He studied at the University of Cambridge and at the University of Sussex, where he obtained his PhD in 1993. He joined the University of Portsmouth in 1996 and was awarded a Royal Society University Research Fellowship in 1999, before being promoted to Professor in the newly formed Institute of Cosmology and Gravitation in 2002.



Paul S. Wesson

University of Waterloo
Department of Physics and Astronomy
Waterloo, Ontario, Canada
psw.papers@yahoo.ca

Chapter E.28

PhD Cambridge, UK, 1979; D.SC. (higher doctorate), London, UK; 2003, F.R.A.S. Research areas: gravitation, cosmology, theoretical astrophysics. Author of several hundred papers and a dozen books, including *Space-Time-Matter*, 2nd edition, 2007.



Nick M. J. Woodhouse

Chapter A.3

University of Oxford and Clay Mathematics Institute
Mathematical Institute, Andrew Wiles Building, Radcliffe Observatory Quarter
Oxford, UK
nick.woodhouse@maths.ox.ac.uk

Nick Woodhouse is a Senior Research Fellow at Wadham College, Oxford, and President of the Clay Mathematics Institute. He did his PhD with Felix Pirani at King's College, London and was a post-doc under John Wheeler at Princeton before joining Roger Penrose's research group in Oxford. His research interests focus on physical applications of geometry and, more recently, on self-duality and integrability.

H. Dieter Zeh

Chapter B.10

Universität Heidelberg
Institut für Theoretische Physik
Heidelberg, Germany
zeh@uni-heidelberg.de



Professor em. H. Dieter Zeh studied at the TU Braunschweig and the University of Heidelberg. Thesis on theory of alpha-decay. Work on nuclear structure (spontaneous symmetry breaking) and nuclear synthesis. He was Postdoc and research associate at Caltech and UCSD. Until his retirement he was Lecturer and Professor at the University of Heidelberg. His work focused on foundations of quantum mechanics (decoherence) and the direction of time.

Detailed Contents

Part B Foundational Issues

Part C Spacetime Structure and Mathematics

Detailed Cont.

Part D Confronting Relativity Theories with Observations

Part E General Relativity and the Universe

Part F Spacetime Beyond Einstein

Index

D

E

Index

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Y

Z